NUDGE

Nudging consumers towards energy efficiency through behavioural science

Deliverable 2.3

Final report on the evaluation of nudging interventions through pilot data

Authors: Anne Kesselring, Sabine Pelka, Sabine Preuß (Fraunhofer Institut für System und Innovationsforschung ISI), Merkouris Karaliopoulos, Andreas Chitos (Athens University of Economics and Business), Peter Conradie, Stephanie van Hove, Emma Martens (imec-mict-UGent)

Project Coordinator: Filippos Anagnostopoulos, filippos@ieecp.org

NUDGE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 957012.

Project information

Project Title	Nudging consumers towards energy efficiency through behavioral science
Project Acronym	NUDGE
Project Number	927012
Project dates	September 2020 – November 2023

Legal Notice

About

Efforts to induce energy-friendly behavior from end-users through behavioral interventions are characterized by a lack of customer personalization ("one-size-fits-all interventions"), a partial understanding of how different interventions interact with each other and contrasting evidence about their effectiveness as a result of poor testing under real world conditions.

NUDGE has been conceived to unleash the potential of behavioral interventions for long-lasting energy efficiency behavior changes, paving the way to the generalized use of such interventions as a worthy addition to the policy-making toolbox. We take a mixed approach to the consumer analysis and intervention design with tasks combining surveys and field trials. Firmly rooted in behavioral science methods, we will study individual psychological and contextual variables underlying consumers' behavior to tailor the design of behavioral interventions for them, with a clear bias towards interventions of the nudging type.

The designed interventions are compared against traditional ones in field trials (pilots) in five different EU states, exhibiting striking diversity in terms of innovative energy usage scenarios (e.g., PV production for EV charging, DR for natural gas), demographic and socio-economic variables of the involved populations, mediation platforms for operationalizing the intervention (smart mobile apps, dashboards, web portals, educational material and intergenerational learning practices).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957012.

Project partners

Table of Contents

List of Figures

List of Tables

Executive Summary

Households play a key role in fulfilling the EU's energy efficiency target of reducing energy consumption by at least 11.7 percent in 2030[1]. Key measures are building renovations and energy labeling of products. They focus on guiding investment decisions and the policy design is built on economic profitability in the long run. Their potential to reduce households' energy consumption depends, above all, on tenure status and household budget. By contrast, energy consumption during the operational phase of these goods depends on households' behavior. Hence, there is a need to understand and exploit households' energy consumption behavior during this phase, especially because behavior accumulates over time beyond the moment an investment decision is made. This need was intensified during the European energy crisis in 2022, when policymakers feared energy shortages and many households faced increasing energy bills. At that time, policymakers scrambled to encourage households to save energy on a voluntary basis – primarily for gas, but in a second instance also for electricity. The available policy measures can be broadly classified as measures with a direct financial incentive (such as taxes or subsidies), and those that stimulate behavior without such a direct incentive. Among those without direct incentives, nudges, i.e., subtle changes in the households' choice architecture, have been proposed as a key approach.

Nudging constitutes a behavioral intervention that does not regulate or ban any options in the choice set and does not provide a specific financial incentive in return for behavioral adjustment. Nudges instead consist of subtle changes in the individual's choice architecture that aim to stimulate behavioral change. Whereas other common stimuli, such as financial incentives, affect the individual's rational reasoning, nudges stimulate the intuitive choices of our everyday life. Crucially, the argument is that nudges can make consumers better off "as judged by themselves". They thus improve consumer satisfaction while contributing to aggregate energy savings as a public good. With the development of smart meter infrastructure and digital applications for the delivery of nudging interventions, there is both practical and scientific interest in understanding how effective these tools can be in shaping behavior.

The project NUDGE was set up to analyze households' behavior regarding energy-saving decisions, and to design and test nudging interventions in residential households across five EU Member States. The core of the project is field experiments at five pilot sites in Germany, Croatia, Portugal, Greece, and Belgium. These experiments targeted energy savings in various contexts, such as heating demand, electricity consumption, and self-consumption for prosumers. Figure 1-1 provides an overview of the five sites, their objectives, and the sample of participants.

The primary goal of **the five field experiments** was to evaluate the effectiveness of different sets of nudging interventions. To this end, we collected different types of data over an 18-month period, comprising survey data, logs of the interaction of end users with mobile apps and online tools, and smart meter data. Since each pilot implements different interventions over different populations, we get a comprehensive set of insights into nudging effectiveness from 13 different nudges in five Member States.

The primary goal of the work reported in this document is to statistically evaluate the effectiveness of nudging interventions implemented in **the five field experiments.**This deliverable focuses on the analysis

[1] https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-targets_en (last visited: 6/12/2023)

of smart meter data, but uses the other data sources to support and probe the main findings. For further analysis of the survey data and the app logs, deliverable D1.3 contains the additional results. Figure 1-1 gives an overview of the field experiments. The underlying hypotheses relate to four outcomes: decreasing (a) electricity consumption and (b) heating demand, as well as increasing (c) prosumer self-consumption, (d) energy knowledge.

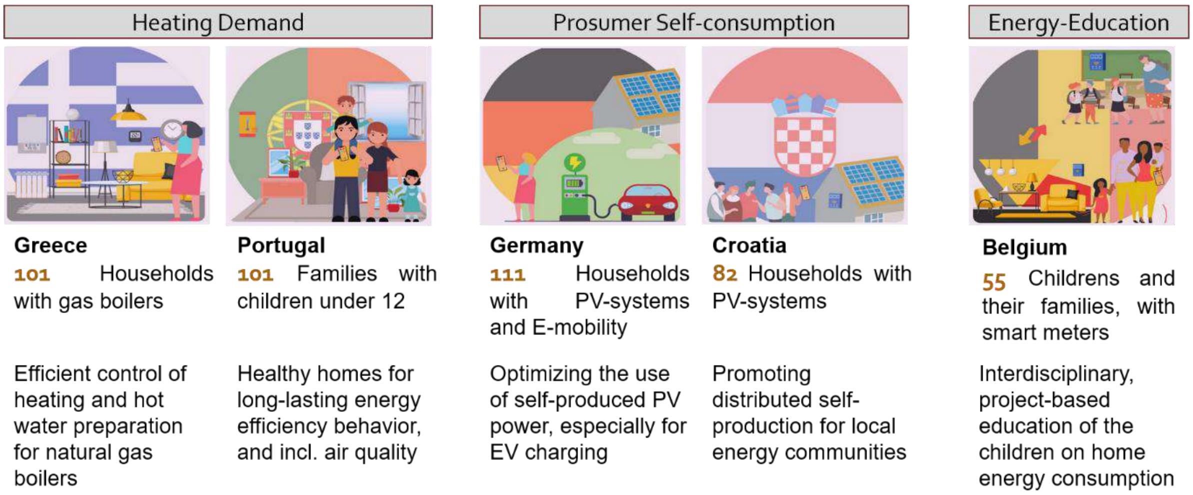


Figure 1-1: Overview of the five pilots: sample description and targeted outcomes

Smart meter data from each field experiment were collected over a baseline period, three intervention periods during which the nudges were administered, and a post-intervention period. The overall timeline covered the period from fall/winter 2021/2022 until later summer 2023. In addition, we collected four waves of survey data, time-aligned with the baseline and the intervention periods, to capture perceived attitudes, preferences, and self-reported behavior. Logs from the interaction of pilot participants with the digital tools deployed in Greece, Portugal, Croatia, and Germany provide additional insights into their exposure to nudging. In the Belgian case, delivery of the nudges occurred through lessons in school rather than digital tools.

The methodology for extracting the causal effects of nudging relies on advanced econometric techniques. The statistical methods need to address (1) confounding factors that drive the studied outcomes across time and (2) systematic differences across households. Most pilots, therefore, featured a treatment and a control group. We employ a "difference-in-differences" design, which is one of the key approaches to causal evaluation in the micro-econometric toolbox (see Section 2.3). We additionally control for differences across the individual households and even the individual days, using a refined version of the design known in statistics as "two-way fixed effects". This allows us to account for time-related confounders, such as seasonality, and household-related factors, such as different dwelling characteristics. The complexity of the econometric evaluation became even more important due to the energy crisis following the war in Ukraine, which unfolded during our field experiments. To study the effectiveness of nudging in a sound manner, the evaluation had to find statistical means to curb the influence of this broader development.

We tailored the general methodology to the specific design, data structure, and country context of each pilot. In the Belgian and Croatian pilots, the design and/or sample structure did not allow a treatment-control-design; hence, we evaluated these pilots through within-subject approaches. Across all pilots, the choice of evaluation strategy was a focal point of the project because the objective is to study the effectiveness of behavioral interventions within households' everyday lives. Micro data from smart metering require substantially more structure and complexity than laboratory experiments, but this form of evaluation is important to assess the potential of nudging in real-life policy portfolios.

The results provide evidence that nudging is highly context-dependent and its effectiveness differs across the pilots and even across interventions *within* the same pilot. We identify several promising cases for nudges, e.g., to stimulate energy savings. In cases, where we measured a significant positive impact, the energy savings reached up to 4 percent for interactive nudges and up to 16 percent for participants who accepted an automated optimization of their consumption (in our case, optimized charging of electric vehicles). Yet, we find that nudging does not work across all groups and contexts. Overall, nudging did not trigger the desired energy savings across all the different country settings and intervention designs. Table 1-1 summarizes the overall effectiveness of nudging in the five pilots as a summary evaluation. Dark green check marks indicate cases with robust evidence supporting the hypothesis. Light green check marks indicate cases with some positive, but mixed evidence. In cases, where the evidence is weak or contradictory, we placed orange minus symbols. In brief, the Table 1-1 underscores that the evidence does not support a consistent picture for nudging effectiveness.

We identified several **explanatory components** hindering the success of the nudges in some pilots: We found low levels of participants' interaction with the digital nudging tools. There are seasonal influences that appear to block nudging effects, such as holiday seasons or periods where there is little room for further improvement due to weather conditions. However, the weak uptake is also an important lesson in itself: nudging can be effective for individual participants, but the systemic effects hinge upon the preliminary step of generating interest vis-a-vis the intervention. Behavioral change depends first on how effective digital tools are in capturing attention that results in exposure.

Table 1-1: Summary of results from field experiments by outcome and pilot site

Nudges are effective in:	**Briefing:**
Increasing the self-consumption ✓ ✓	• Single digit effect sizes in Germany, but with stronger increases for the smart charging nudge • Mixed results in Croatia attributed to regulatory setting
Reducing the electricity consumption ✓ –	• Decrease in household consumption in Germany • Mixed results including positive and negative coefficients as well as group divergence in Portugal
Reducing the heat consumption ✓ –	• Reductions observed only in one sub-group for Belgium • Weak/contradicting evidence in Greece, with comparatively low level of activity on the mobile app
Improving the energy knowledge level –	• No evidence for improvement in test scores, with concerns about comparability in test results

Our results bring forth several **recommendations**. Above all, scaling up nudging to broader populations must consider the pre-conditions of behavioral change carefully. Acknowledging the prevalence of low interaction, we recommend that service providers and policymakers choose nudges with an immediate response (e.g., defaults) where possible. This is especially important considering that our participants are likely above average in (a) motivation and (b) means to respond. In the design and implementation of nudging interventions, our results thus identify constraints that must be considered to gain a realistic expectation of nudging in policy portfolios. Moreover, we see a need to better align the nudging intervention with the overarching frameworks and incentives from a regulatory and market side. The prime example in our experiment is the Croatian regulatory system, which penalizes self-generation by rooftop-photovoltaic that exceeds the level of household consumption. Thus, in the pilot, the participants use digital tools to monitor compliance rather than respond to nudges. This example of a regulatory framework offsetting the nudging intervention shows that if the hindering framework conditions cannot be refined, the nudges must be adapted in scope and content. In particular, they must be carefully aligned to the existing framework conditions.

The findings regarding the conditional effectiveness of nudging are in line with existing insights from the literature. We contribute new evidence to previous literature with our extensive data basis, research design with control groups, and applied causal-effect methods that allow us to study nudging within the context of participants' everyday lives.

Previous literature on nudging had pointed to many of the factors we discuss as limitations, but the studies applied for individual cases or designs that were difficult to compare. Overall, the literature is fragmented, with high promise and disappointing results scattered across lab and field experiments.[2] In particular, the academic literature may present an overly optimistic picture of nudging effectiveness due to publication bias.[3] In addition, nudging experiments in controlled settings (e.g., labs, surveys) may not be scalable in real-life settings.[4,5] Our results combine pieces of evidence that together deliver a wide scope of nudging interventions developed from a common base and discussed across pilots over the entire period. With this effort, the findings from the project NUDGE support a cautious view on nudging as a tool that cannot be used as a blanket measure, even with digital tools and easy access to information. Instead, targeting and tailoring to local contexts are critical pre-conditions to successful implementation.

The main message from the project experiments is that nudging is a soft intervention that is easily dominated by external circumstances. Our real-life field trials revealed that several factors overlay everyday decisions and, therefore, limit the effectiveness of nudging. Overall, the results indicate that nudging is more effective in selected circumstances where there is a direct link between the treatment and the outcome. Moreover, it is important that the timing is suitable for guiding behavioral change. When those restrictive preconditions are not met, the nudges do not cut through the host of other factors that shape energy consumption by private households. The field experiment contributes to the understanding of nudging by tracking behavior with extensive data coverage over time and across countries. Using advanced econometric techniques and high-frequency smart meter data, the results provide deeper insights into the

[2] Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. Proceedings of the National Academy of Sciences, 119(1), e2107346118.

[3] Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J., & Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e2200300119.

[4] Allcott, H., & Taubinsky, D. (2015). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, *105*(8), 2501-2538.

[5] Alberini, A. (2019). Revealed versus stated preferences: what have we learned about valuation and behavior?. *Review of Environmental Economics and Policy*.

channel of nudging from intention/ motivation to behavioral changes measured through objective consumption data. Tracking households over a period of more than 1.5 years allows for deeper insights into the effects of nudging in the context of everyday decisions, where incentives and exposure naturally change throughout the year. Collectively, the results emphasize that the question of whether nudging is effective is insufficient, as the underlying “where, how, and why can nudging be effective” is needed to get a better sense of the potential of behavioral interventions.

NUDGE
Nudging consumers towards energy efficiency through behavioural science

1 Introduction

The EU targets decreasing final energy consumption by at least 11.7 percent by 2030[6]. Households represent a large share of energy consumption in the EU. In 2021, they accounted for 27 percent of final energy consumption in the EU. Thereof, natural gas covered 33.5 percent of the households' energy consumption[7]. Thus, reducing households' final energy demand is key for achieving the EU's energy efficiency targets. The European energy crisis in 2022 with gas shortages and increasing energy bills intensified this need for action. Key actions on this energy efficiency pathway are renovations of buildings and energy labeling of products. Yet, there is a need for additional actions facilitating immediate energy savings for all households - especially when they are under pressure of increasing energy bills.

In general, households base their consumption on intuitive decisions (e.g., switching on and off heating or lights, using white goods). To stimulate changes in their consumption decisions, consumption information, e.g., provided via smart meters or smart home systems, can be used. Yet, this information needs to be provided in a special way responding to these intuitive decisions. Such subtle changes in how information is provided and decisions are made are called nudges.

Nudges are any change of the decision context "that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentive"[8]. They are choice-preserving and do not address the individuals' rational reasoning. Unlike traditional policy interventions, they do not fundamentally change the rules of the game but support individuals to better play the game[9]. Nudges comprise even changing the decision context that does not affect the pre-defined choice (i.e., defaults). Thereby, nudges target different degrees of re-evaluation during the individuals' intuitive decision-making. Lin et al. 2017 distinguish between two types of changes. Firstly, they identify changes of the immediate decision context, e.g., just-in-time prompt, when increasing the heating temperature. Secondly, they identify changes in the indirect decision context, such as comparisons to the national average consumption on the heating bill[10]. In the case of the latter, the treatment and its effect on behavior spread over time.

Existing literature discusses controversially which nudges are most effective for which kind of group and context. Social comparisons, feedback, and using defaults are the most popular nudges in energy-related literature.

---

[6] https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-targets_en

[7] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households

[8] R. H. Thaler, C. R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Yale University Press, 2008).

[9] Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences, 46*, E147. doi:10.1017/S0140525X22002023

[10] Yiling Lin, Magda Osman & Richard Ashcroft (2017): Nudge: Concept, Effectiveness, and Ethics, Basic and Applied Social Psychology, DOI: 10.1080/01973533.2017.1356304

Nudges employing *social comparison* incite energy savings playfully and competitively. They are the most researched[11] and are identified in some studies as the most promising ones[12]. They are especially suitable for reinforcing existing sustainable behavior[13]. Social comparison studies report that nudges employing social comparison can lead to 1.2 to 30 percent energy savings[10]. The American service company Opower conducted a set of studies producing savings on the lower end of the spectrum. Due to long observation periods and large samples, they can be considered robust saving projections[10].

Smart metering studies most frequently apply *feedback* on consumption behavior and its consequences. The processed meter data is suitable to increase the individual's knowledge [14] and avoid unsustainable behavior[11].On smartphone apps, push notifications or just-in-time prompts, which pop up during an energy-related action (e.g., adjusting the room temperature) sometimes accompany the feedback. Feedback studies report 5 to 13 percentage of energy savings[15,16,17,18,19,20,21]. The efficacy of feedback depends on the interest level of the participants, the appropriate timing and frequency of the feedback, as well as the message framing[13]. Messages addressing cognitive biases (e.g., loss aversion[18]) or multiple benefits (e.g.,

---

[11] Andor MA, Fels KM. Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects. Ecological Economics 2018;148:178–210. https://doi.org/10.1016/j.ecolecon.2018.01.018.

[12] Bergquist M, Thiel M, Goldberg MH, van der Linden S. Field interventions for climate change mitigation behaviors: A second-order meta-analysis. Sustainability Science 2022.

[13] Beermann, V.; Rieder, A.; Uebernickel, F. (2022): Green nudges: how to induce. Pro-Environmental Behavior Using Technology. Forty-Third International Conference on Information Systems, Copenhagen 2022.

[14] Abrahamse, Wokje; Steg, Linda; Vlek, Charles; Rothengatter, Talib (2005): A review of intervention studies aimed at household energy conservation. In: *Journal of Environmental Psychology* 25 (3), S. 273–291. DOI: 10.1016/j.jenvp.2005.08.002.

[15] Houde S, Todd A, Sudarshan A, Carrie Armel K. Real-time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence. EJ 2013;34(1). https://doi.org/10.5547/01956574.34.1.4.

[16] Asensio OI, Delmas MA. Nonprice incentives and energy conservation. Proc Natl Acad Sci U S A 2015;112(6):E515. https://doi.org/10.1073/pnas.1401880112

[17] Ruokamo E, Meriläinen T, Karhinen S, Räihä J, Suur-Uski P, Timonen L et al. The effect of information nudges on energy saving: Observations from a randomized field experiment in Finland. Energy Policy 2022;161:112731. https://doi.org/10.1016/j.enpol.2021.112731.

[18] Dominicis S de, Sokoloski R, Jaeger CM, Schultz PW. Making the smart meter social promotes long-term energy conservation. Palgrave Commun 2019;5(1). https://doi.org/10.1057/s41599-019-0254-5.

[19] Bager S, Mundaca L. Making 'Smart Meters' smarter? Insights from a behavioral economics pilot field experiment in Copenhagen, Denmark. Energy Research & Social Science 2017;28:68–76. https://doi.org/10.1016/j.erss.2017.04.008.

[20] Myers E, Souza M. Social comparison nudges without monetary incentives: Evidence from home energy reports. Journal of Environmental Economics and Management 2020;101:102315. https://doi.org/10.1016/j.jeem.2020.102315.

[21] Schleich J, Faure C, Klobasa M. Persistence of the effects of providing feedback alongside smart metering devices on household electricity demand. Energy Policy 2017;107:225–33. https://doi.org/10.1016/j.enpol.2017.05.002.

reduction of air pollution[15]) are more effective than neutrally framed ones. Framing financial implications shows ambiguous results. It led to reinforcing[22] or attenuating effects[23]. Research designs with feedback without financial implications (e.g., energy savings in college dorms) lead to no effect at all[19]. Feedback mediums with low (e.g., monthly energy report) and high interaction (e.g., indoor displays), were reported to be equally effective[17], whereas home indoor displays showed a high efficacy[10].

Studies *combining feedback and social comparisons* show ambiguous results. In a 2021 study, low levels of observation led to no effects for social comparison [13]. By contrast, a study in 2019 showed only a positive effect, when the social comparison was added to the feedback[14].

Using *defaults* is suitable for creating new routines for operating emerging innovations[12]. They help individuals make decisions in new situations, which are subject to uncertainty. Instead of being inert and loss-averse, the individuals are released from their decision-making by the preset defaults[24]. A study in 2014 reported a remarkable 44.6 percent increase in adopting green electricity tariffs by setting them as default tariffs[25].

In our study, we test in five pilots how nudges can support households in changing their behavior. As studies with combined nudges showed promising results[10], we selected the most promising nudges, feedback, and comparison. We combine them in an accumulative manner, which allows the evaluation of the incremental effect of single nudges. However, the breadth of findings in the literature is wide:

- Feedback delivered via digital tools can increase awareness on their consumption behavior.
- Just-in-time prompts and push notifications can stimulate immediate responses from the participants.
- Ambient feedback makes the participants internalize energy savings practices and creates an intuitive dissonance if the participants do not comply with it (e.g., by green-coloring the dashboard items during low electricity consumption).
- Empathy-instigating messages (e.g., pictures of a polluted earth), gamification and target setting aim to stimulate behavioral change in a playful or emotional manner.
- Defaults and opt-in nudges bypass the individuals' inertia of decision-making but preserve the individuals' right to choose.

Thus, it is difficult to identify the most promising nudges for energy savings and allocate them to specific households and contexts. Moreover, the context-specific nature of nudging studies complicates

---

[22] Julian Huber, Dominik Jung, Elisabeth Schaule, and Christof Weinhardt: GOAL FRAMING IN SMART CHARGING - INCREASING BEV USERS' CHARGING FLEXIBILITY WITH DIGITAL NUDGES, Conference: 27th European Conference on Information Systems (ECIS), 2020

[23] Sudarshan, Anant (2017): Nudges in the marketplace: The response of household electricity consumption to information and monetary incentives. In: Journal of Economic Behavior & Organization 134, S. 320–335. DOI: 10.1016/j.jebo.2016.12.015

[24] Schubert C. Green nudges: Do they work? Are they ethical? Ecological Economics 2017;132:329–42. https://doi.org/10.1016/j.ecolecon.2016.11.009.

[25] Momsen K, Stoerk T. From intention to action: Can nudges help consumers to choose renewable energy? Energy Policy 2014;74:376–82. https://doi.org/10.1016/j.enpol.2014.07.008.

comparisons between different nudge interventions. We address this challenge along three context dimensions. First, we implement three different nudges in the same pilot setting to make their nudging effects comparable to each other. Second, we test the nudges for three energy-efficient behaviors: decreasing electricity consumption, decreasing heating consumption, and increasing self-consumption (see Table 1-1). We also test two further energy-related behaviors, improving energy knowledge and indoor air quality. Third, we test changes in each targeted energy-efficient behavior in at least two pilot sites. Thereby, we can identify national differences and similarities. While almost all pilots target the electricity consumption of their participants, the Greek, Belgian, and Portuguese pilots also target their heating consumption. The German and Croatian pilots engage participants with rooftop-photovoltaic and aim to increase their self-consumption.

Table 1-1: Consumption behavior addressed in each pilot site and respective structuring of the current deliverable into sections

Targeted consumption behavior	Germany (Sec. 3)	Croatia (Sec. 4)	Belgium (Sec. 5)	Portugal (Sec. 6)	Greece (Sec. 7)
Increasing self-consumption	X	X			
Decreasing heating consumption			X	X	X
Decreasing electricity consumption	X			X	
Enhancing energy knowledge			X		
Improving indoor air quality				X	

The comparative assessment of the different nudges within one pilot and the differences between pilots targeting the same energy-efficient behavior and between the ones with different behaviors enable multiple insights into the effectiveness of nudges in different contexts. Thereby, we analyze the effect of the nudges on the actual consumption behavior based on sensor data, as well as how the participants perceived their behavior based on survey data. Contrasting both data sources reveals drivers and barriers of behavioral change and identifies participants' intention-action gaps.

After introducing general methodological aspects of our analysis in Section 2, sections 3-7 present the nudging effects for each pilot site, as these are estimated based on sensor and survey data. These sections are ordered in line with the grouping of pilot sites according to the consumption behavior that the nudges address, as shown in Table 1-1. In Section 8, we synthesize the pilot-specific findings.

2 Methodology

The following section outlines materials and methods to evaluate the nudge interventions. It contains general aspects later specified in more detail for each pilot in the following pilot-specific sections.

2.1 Experiment Design

Three nudging interventions are sequentially tested in each of the four pilots, excluding the Belgian one that delivers its educational nudges through courses during the school year. The experiment design for each pilot was selected based on the research methodology presented in Deliverable 2.2[26].

The collection of data over time (time series) and across participants (cross-sectional) allows us to design the experiments in a way that lets us analyze consumption changes over time, between the participants, or changes between the participants over time. The latter, i.e., comparing the changes over both dimensions (time and participants), reduces the impact of confounding factors (see Section 2.4). It requires longer nudging periods and larger samples, which can be divided into control and treatment groups. We confirmed these criteria for the German, Croatian, and Portuguese pilots. We divided their samples randomly into two equally large groups and treated them in an alternating manner: in the first half of a nudging period, the first group serves as the treatment group, and the second group as the control group. In the second half of the nudging period, we switch the assignments for both groups, i.e., the second group is treated. Since both groups are treated during the first or second half of the nudging period, we can determine a treatment effect for both groups.

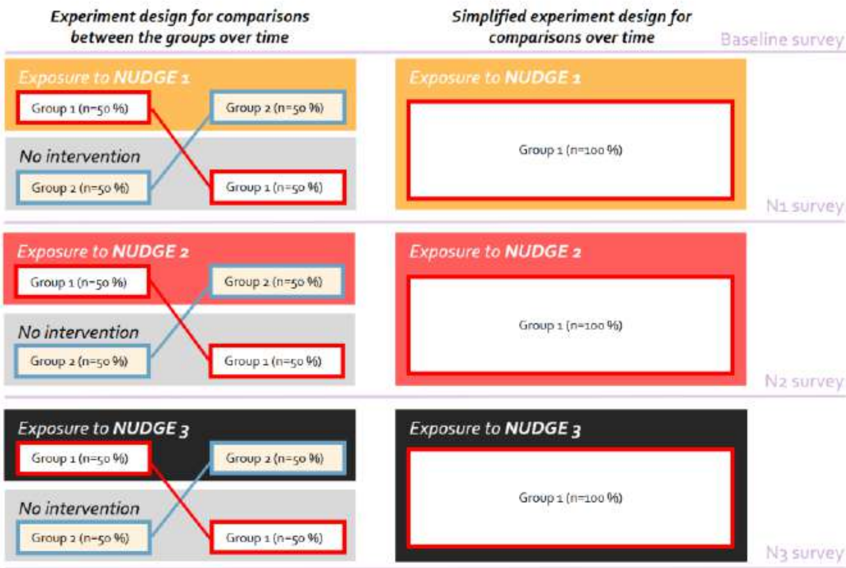


Figure 2-1: Experiment design for the pilots in Germany, Portugal and Croatia (left side) and in Greece and Belgium (right side)

26 https://www.nudgeproject.eu/wp-content/uploads/2022/08/D2.2-Research-methodology-for-assessing-the-effectiveness-of-interventions-regarding-change-of-energy-efficient-behavior.pdf (last visited:21/12/2023)

For the Belgian and Greek pilots, the expected amount of collected data over time and across the participants did not allow this split of the sample and nudging period without violating the power constraints of the analysis. In particular, the Belgian pilot is characterized by a smaller sample of 16 (cohort 1) and 25 pupils (cohort 2), and the Greek pilot about gas consumption for heating purposes is limited to the heating season. For these two pilots, all participants were treated during the entire nudging period.

When selecting the nudging interventions for each pilot, we consider the targeted consumption behavior, the technical conditions at the pilot site, and potential learning or fatigue effects. Most pilots started with nudges that provided feedback on the participants' consumption and aimed to increase their awareness (see Table 2-1). More interactive nudges, in particular ones with push notifications, just-in-time-prompts, gamification or target settings, followed these. The specific nudge design is introduced in the pilot-specific nudge sections.

Table 2-1: The three nudging interventions realized in each of the five NUDGE pilots

	Germany (Section 3)	Croatia (Section 4)	Belgium (Section 5)	Portugal (Section 5)	Greece (Section 6)
Nudge 1	Feedback & awareness	Instigating empathy	Energy courses & pupils as multipliers for two school cohorts	Feedback & awareness	Feedback & awareness
Nudge 2	Gamification & target setting	Feedback & awareness		Push-notifications	Just-in-time prompts
Nudge 3	Opt-in	Gamification & target setting		Push-notifications, feedback & awareness	Push-notifications

2.2 Data

We evaluate the participants' perceived effect based on the collected survey data and their actual effect based on the collected sensor data. The survey data were collected before the first and after each nudging period (in total four times, see Section 2.2). The sensor data were continuously collected during the nudging periods, but also before and after them. Meta-data on the weather, the household composition, and the interaction with the nudging tools are available to assess the corresponding factors' impact on the nudging effects.

Three consumption behaviors and energy-related behaviors are targeted in the pilots. Most pilots have multiple targets, either pursued simultaneously or sequentially (e.g., Portuguese pilot). Heating consumption aimed to be decreased in the Belgian, Greek, and Portuguese pilots. While the Belgian and Greek pilots focused on the gas consumption of their participants, the Portuguese ones monitored the electricity consumption for heating. German and Croatian participants own rooftop-photovoltaic and aim to increase the consumption of their self-generated electricity (i.e., their self-consumption). We also determine the share of self-consumption relative to their overall consumption, which is called autarky rate. All pilots, apart from the Greek and Croatian ones, also aim to decrease their electricity consumption. How these key performance indicators (KPIs) are measured and calculated is described in Deliverables 1.2 and 4.1 [27].

The Portuguese and Belgian pilots also aim to stimulate other energy-related behaviors besides energy savings. In the Portuguese pilot, the goal is to improve indoor air quality (IAQ). Its participants are equipped with IAQ sensors to monitor the concentration of particles in the air. In the Belgian pilot, the goal is to enhance awareness about energy matters among the pupils themselves but also their families and the way to monitor this is through a quiz embedded into the survey.

The corresponding hypotheses for the KPIs of Table 2-2 are introduced in the pilot-specific sections.

Table 2-2: Sensor data-based KPIs for assessing the effect of nudging interventions in each of the five NUDGE pilots

<table>
<tr><th>Targeted consumption behavior-key performance indicator (KPI)</th><th>Germany (Sec. 3)</th><th>Croatia (Sec. 4)</th><th>Belgium (Sec.5)</th><th>Portugal (Sec. 6)</th><th>Greece (Sec. 7)</th></tr>
<tr><td>Increasing self-consumption</td><td colspan="2">Self-consumption [Wh], autarky rate [%]</td><td></td><td></td><td></td></tr>
<tr><td>Enhancing energy knowledge</td><td></td><td></td><td>survey-based</td><td></td><td></td></tr>
<tr><td>Decreasing heating consumption</td><td></td><td></td><td>Gas consumption [kWh]</td><td>Electricity consumption [kWh][1]</td><td>Gas consumption [kWh]</td></tr>
<tr><td>Decreasing electricity consumption</td><td colspan="4">Electricity consumption [kWh]</td><td></td></tr>
<tr><td>Improving indoor air quality (IAQ)</td><td></td><td></td><td></td><td>IAQ [e.g., CO_2][1]</td><td></td></tr>
<tr><td colspan="6">1 This deliverable focuses on decreasing the electricity consumption for the Portuguese pilot. Other KPIs are evaluated in Deliverable 1.3</td></tr>
</table>

For the self-reflection on consumption behavior, we asked the participants for their intention and motivation to save energy in the surveys, in particular before the first and after each nudging intervention.

---

[27] While Deliverable 1.2 is confidential, Deliverable 4.1 can be found in the NUDGE knowledge hub: https://www.nudgeproject.eu/knowledge-hub/

Apart from the generic target of saving energy, we added intention items with the specific target of the pilots, e.g., increasing the self-consumption for the German and Croatian pilots. For comparability across pilots, we focus our analyzes on the generic energy saving-items. For the motivation items, we used the METUX model by Peters, Calvo, and Ryan (2018)[28] and its items, ranging from extrinsic to intrinsic motivation. The items (see Table 2-3) are measured on a 5-point Likert scale, ranging from 1 – strongly disagree to 5 – strongly agree.

Table 2-3: Survey items for measuring motivation and intention

Psychological aspect	Item
Motivation	I expected it will be interesting to save energy
	It will help me do something important to me
	I think it would be enjoyable
	It is going to be of value to me in my life
	I want others to know I save energy
	I am required to save energy (e.g., by a research study)
	I feel pressured to save energy
	It will look good to others if I save energy
Intention	I tried to save energy at home in the last [duration of nudging period] months.
	I think I have saved energy at home in the last [duration of nudging period] months.
	I will try to save more energy at home in the next [duration of nudging period] months.

We expect the nudges to increase the intention and motivation for saving energy. This leads to the following general hypotheses:

All1: Nudges are effective in increasing the intention to save energy of participants.

All2: Nudges are effective in increasing the motivation to save energy of participants.

The sample sizes of the individual survey waves should ideally be identical to the overall sample size of the pilot. Still, there were slight differences in sample size due to an imperfect response rate and the subsequent exclusion of outliers. The increasing sample size over the waves is striking for the Croatian pilot. This is because while Nudge 1 had already been carried out, other people registered to participate. We illustrate the socio-demographical data of each wave in Table 2-4.

[28] Peters, D.; Calvo, R. A.; Ryan, R. M. (2018): Designing for motivation, engagement and wellbeing in digital experience. Front. Psychol. 9: 797.10.3389/fpsyg.2018.00797.

Table 2-4: Socio-demographical data of each survey wave

	Baseline (wave 1)	Nudge 1 (wave 2)	Nudge 2 (wave 3)	Nudge 3 (wave 4)
German pilot				
Sample (outlier excluded)	99 (6)	86 (4)	91 (4)	85 (2)
Stated being female	12	10	10	9
Stated being male	87	76	78	74
Average age (SD)	56.57 (10.22)	56.93 (10.31)	56.57 (10.67)	56.28 (10.49)
Average apartment size in m² (SD)	167.56 (55.11)	167.09 (55.09)	166.3 (54.46)	159.85 (50.08)
Number of persons in household	3 (2.90)	3 (2.85)	3 (2.90)	3 (2.82)
Average days per week being mainly at home	5.69	5.7	5.45	5.58
Sub-sample group 1	47	20	44	41
Sub-sample group 2	35	23	34	44
Croatian pilot				
Sample (outlier excluded)	82 (0)	43 (7)	78 (0)	77 (3)
Stated being female	6	4	6	6
Stated being male	76	39	72	71
Average age (SD)	47.12 (9.16)	47.67 (8.76)	47.43 (9.28)	47.45 (9.33)
Average apparment size in m² (SD)	188.17 (87.7)	172.42 (62.99)	187.56 (84.46)	184.42 (86.06)
Number of persons in household	4 (4.29)	4 (4.28)	4 (4.29)	4 (4.26)

Average days per week being mainly at home	5.87	5.93	4.41	4.52
Sub-sample group 1	47	20	44	44
Sub-sample group 2	35	23	34	33
Belgian pilot	***Cohort 1 - Experiment***	***Cohort 1 - Control***	***Cohort 2 - Experiment***	***Cohort 2 - Control***
Sample (outlier excluded)	30	22	32	23
Stated being female	10	14	16	16
Stated being male	20	8	16	7
Average age (SD)	43.27 (5.49)	41.41 (3.51)	41.75 (3.23)	40.22 (4.52)
Average apartment size in m² (SD)	225.7 (104.96)	253.82 (93.96)	216.28 (133.21)	228 (188.86)
Number of persons in household	3.9	4.00	3.75	3.91
Average days per week being mainly at home	5.57	5 .39	4.17	4. 70
Portuguese pilot	**Baseline (wave 1)**	**Nudge 1 (wave 2)**	**Nudge 2 (wave 3)**	**Nudge 3 (wave 4)**
Sample (outlier excluded)	86 (0)	65 (6)	62 (8)	72 (10)
Stated being female	42	31	30	34
Stated being male	44	34	32	38
Average age (SD)	39.89 (6.59)	40.17 (6.48)	40.87 (6.94)	40.69 (6.64)
Average apartment size in m² (SD)	185.55 (264.90)	188.05 (290.39)	196.79 (296.32)	199.33 (286.36)
Number of persons in household	4 (3.84)	4 (3.66)	4 (3.76)	4 (3.81)

Average days per week being mainly at home	6.57	4.75	4.87	4.79
Sub-sample group 1	43	31	32	36
Sub-sample group 2	43	34	30	36
Greek pilot				
Sample (outlier excluded)	41 (19)	37 (1)	81 (4)	72 (1)
Stated being female	8	7	22	20
Stated being male	33	30	58	51
Average age (SD)	34.83 (9.43)	34.54 (9.87)	36.48 (10.71)	36.51 (10.96)
Average apartment size in m² (SD)	82.32 (24.90)	78.65 (23.29)	85.80 (38.38)	80.83 (29.88)
Number of persons in household	2 (1.28)	2 (2.38)	2 (2.15)	2 (2.18)
Average days per week being mainly at home	3.00	5.37	5.65	5.30

2.3 Analytical Strategy

We estimate the actual effect of the nudging interventions based on sensor data and the perceived effect based on survey data. We synthesize both perspectives in the summary chapter of each pilot-specific section. Additionally, we conduct robustness checks of the sensor data analysis based on the survey data and app data (e.g., limiting the analyzed sample to participants using the app, or reporting to be doing so regularly).

Two evaluation strategies exist for the assessment of the nudges: comparing the differences in consumption over time or between the treated and non-treated participants. When the difference in both dimensions is used, the econometric technique is called Difference-in-Differences (DiD). This strategy absorbs the effect of factors that vary over time for *all* participants (e.g., weather) or that vary between participants in *all time steps* (e.g., equipment of the house) and allows us to determine the causal effect of only the nudges on consumption behavior. In other words, when there are similar patterns (*parallel trends*) in the consumption of the treatment and control groups in the period before the treatment, the baseline, any deviations from this common trend across groups during the treatment period can be safely attributed to the nudge (see Figure 2-2). The DiD analysis is implemented with a regression model that estimates the "treatment effect" as shown below. The resulting number answers the question: did the treatment group change *more* than the control group between the baseline period and the nudge period?

The method is considerably more complex and data-demanding than a simple pooled ordinary least squares regression, and it critically hinges on the parallel trends assumption. However, the strategy has important advantages in real-life field experiments. The difference-in-differences strategy addresses more fundamental differences over time by comparing the treatment group to the control group, while also accounting for group differences that persist over time, i.e., that were already there before any intervention. It, therefore, combines the cross-sectional and time-series dimensions of panel data to account for both sources of variation that each could be a confounder to extracting the effect of the nudge from the broader patterns in the data.

When individuals make decisions in their everyday life, there are natural developments over time and some fundamental differences between the groups that could confound the effects. In a lab experiment or a survey, these are less critical because the setting contains less noise and a clear channel. But when studying everyday intuitive decisions, nudging intervention happens within the context. This makes it much harder to obtain credible results, but it allows us to study the effectiveness of nudging in real-life settings, where it is difficult to rule out confounders based on conceptual arguments or controlled environments.

Recent meta-studies highlight the need for selecting methods for testing causal effects[29] [30] and considering the context-specific efficacy of nudges[31]. Regarding the first aspect, suitable methods, such as DiD, reached a level of maturity[32] and are more frequently applied for smart metering studies[33] [34] [35] [36]. Regarding the latter, applying sequentially a set of nudges to a larger, more heterogeneous sample in the same context allows us to compare the efficacy between the nudges and the sub-groups. At the same time, the composition of a

[29] Bergquist M, Thiel M, Goldberg MH, van der Linden S. Field interventions for climate change mitigation behaviors: A second-order meta-analysis. Sustainability Science 2022.

[30] Andor MA, Fels KM. Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects. Ecological Economics 2018;148:178–210. https://doi.org/10.1016/j.ecolecon.2018.01.018.

[31] Lehner M, Mont O, Heiskanen E. Nudging – A promising tool for sustainable consumption behaviour? Journal of Cleaner Production 2016;134:166–77. https://doi.org/10.1016/j.jclepro.2015.11.086 .

[32] Imbens GW, Wooldridge JM. Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature 2009;47(1):5–86. https://doi.org/10.1257/jel.47.1.5 .

[33] Bager S, Mundaca L. Making 'Smart Meters' smarter? Insights from a behavioural economics pilot field experiment in Copenhagen, Denmark. Energy Research & Social Science 2017;28:68–76. https://doi.org/10.1016/j.erss.2017.04.008

[34] Brown Z, Johnstone N, Haščič I, Vong L, Barascud F. Testing the effect of defaults on the thermostat settings of OECD employees. Energy Economics 2013;39:128–34. https://doi.org/10.1016/j.eneco.2013.04.011.

[35] Schleich J, Schuler J, Pfaff M, Frank R. Do green electricity tariffs increase household electricity consumption? Applied Economics 2022:1–12. https://doi.org/10.1080/00036846.2022.2102574.

[36] Myers E, Souza M. Social comparison nudges without monetary incentives: Evidence from home energy reports. Journal of Environmental Economics and Management 2020;101:102315. https://doi.org/10.1016/j.jeem.2020.102315.

larger sample tends to involve more uninterested participants. This leads to a higher external validity to the detriment of a lower effect size [30].

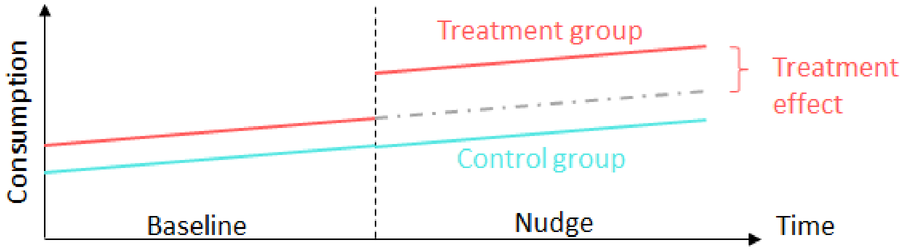


Figure 2-2: Schematic representation of the difference-in-differences method

The DiD approach requires a control group. In the case of a switch of the control and treatment assignment to the groups, it also requires at least one measurement after each switch, and, ideally, multiple observations in each period. The latter is only the case for the continuously measured sensor data. A within-subject comparison is conducted based on the survey data.

We apply the DiD approach to the German, Croatian, Portuguese, and Greek pilots for the sensor data analysis. The alignment between the comparison approach and experiment design involves two deviations from the original experiment design. First, the missing control group in the Greek pilot was replaced with non-active participants. Non-active participants are those who are never using the mobile app and, therefore, have no exposure to the nudging interventions throughout the study period. Second, a hampered participant acquisition and shortages in sensor supply resulted in a successively increasing sample size in the Croatian pilot. The underpowered baseline period limits the robustness of a DiD approach. Therefore, we also conducted a within-subject comparison for the Croatian pilot. Finally, the small sample of the Belgium pilot also results in a within-subject comparison, but this was expected from the design stage (see Table 2-5). In essence, our analytical strategy is to employ the advanced evaluation where possible because we believe that it is important for the credibility of our results, especially considering the pandemic and the energy crisis that re-enforced existing concerns about confounding factors. Nevertheless, we examined the data of each pilot after collection to assess the feasibility and adjust where necessary. The general principle is that the method follows the data to give the best possible answer regarding the effectiveness of nudging in each pilot.

Table 2-5: Assessment methodology for nudging intervention effects in the five pilots

Effect assessment strategy	Germany (Section 3)	Croatia (Section 4)	Belgium (Section 5)	Portugal (Section 5)	Greece (Section 6)
Sensor data: Difference-in-differences	X	X		X	X
Sensor data: Within-subjects		X	X		
Survey data: Within-subjects			X		

A survey testing the intention and motivation to save energy was conducted before the first nudge (baseline) and after each nudge. For the statistical comparison of the two intention values and the motivation value before and after a nudging intervention (between different time points; within subjects), t-tests for dependent samples were used. We excluded outliers with extreme values (M +/- 3SD) from the t-test analyzes.

2.4 Challenges, Analytical Decisions & Risk Estimation

Field trials with pilots allow researchers to observe the decisions of households in their natural everyday life environment. Still, challenges exist that need to be addressed by analytical decisions and a risk assessment strategy. This involves dealing with confounding factors, learning effects, and lower than expected data quality.

We explore the way in which the nudge effect works by examining its time- and group-dependent effects. Time-dependent effects may reveal fatigue in the participants' engagement. For the group-dependent effects, we distinguish between the participants' location, their equipment or their engagement with the digital tools. Furthermore, deep dives further investigate surprising findings.

2.5 Impact of confounding factors

The experimental setup and exogenous factors can influence the impact of the nudges. On the one hand, randomizing a sample of a limited size, such as in our case, might lead to unbalanced sub-groups concerning their household equipment, routine, and socio-demographics. For instance, participants with more installed photovoltaic capacity have more leeway to increase their self-consumption than the ones with less installed capacity. On the other hand, changes over time, such as weather or energy prices, influence household decisions.

The panel data structure of the sensor data allows us to control time- and subject-dependent confounding factors by implementing two-way-fixed effects (TWFE). We use fixed effects for households and days, which is explained in more detail below. It is a powerful, state-of-the-art econometric approach used widely in settings where individuals are very different from each other, and overall conditions change over time. This is the case for our pilots. The first layer of fixed effects cancels out differences across households. Each household has its own fixed effect (intercept term in the regression equation), which accounts for all factors that are fixed for that household over time, e.g., equipment, family size, and building conditions. The second layer of fixed effects absorbs factors that vary over time but in a way that is common for all households in the treatment and control groups, e.g., weather, and news updates. Using fixed effects at the household and day level, rather than aggregated indicators (e.g., by group and by period) is a conservative choice that we believe is necessary in the context of our pilots, as explained in section 2.3.

At the same time, TWFE absorbs only differences of one dimension if the other dimension is constant. For instance, if households have different heating consumption levels and the heating season creates differences that surface only on cold days, it is challenging to absorb differences in the ownership of heat pumps as subject-dependent factors. If households respond differently to increased energy prices, we cannot absorb their differences as a time-dependent factor.

Another exogenous factor, which is hard to absorb with TWFE, is the different optimization potential depending on the weather. For instance, if the generation of households with rooftop-photovoltaic covers 100% of their midday consumption in summer, there is no leeway to optimize their self-consumption further.

We carefully consider the limitations of the TWFE modeling approach and the impact of exogenous factors in the discussion of the results.

2.6 Learning effects

Testing different nudges with the same sample allows us to compare the nudges' effects in the same setting. At the same time, the treatment with one nudge might affect the behavior of participants in the long run, even though the nudge is removed afterwards (learning effects). Starting with more subtle nudges and continuing with more interactive ones mitigates this risk.

In the case of switches between control and treatment assignment of two groups, the control group of the second half, which was treated before, is no "clean" control group because those participants had already been exposed to the nudge. This is both a limitation and an asset of our approach. From the theory behind nudging, there should not be learning effects because the nudge speaks only to the intuitive system. This assumption has been questioned in empirical studies. Hence, differences between the first and second group can give insights into learning. However, while the project was already underway, new scientific results became available that indicate that the established statistical methods can be problematic when the control group is not "clean."[37] We carefully consider this limitation in the interpretation of results.

2.7 Data availability and quality

The nudge evaluation is based on a rich data set collected from five pilots with 41 to 111 participants over 1.5 years. Issues with the technical equipment or human shortfalls impact the quality of these data. We identified four issues for data quality and three strategies for how to cope with them.

---

[37] For a review of the recently identified issues, see e.g., Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, 144(2), 37395.

First, acquiring participants and keeping them involved throughout the experiment was challenging for the pilot partners, which they managed with high commitment. The dropout rates were low. However, it took longer for the Croatian and Greek pilots to reach their full sample than scheduled. We decided to postpone the start of the first nudge for the Croatian pilot to involve a larger number of participants. Since the heating period is the only eligible intervention phase for the Greek pilot, postponing the start of the nudge for a larger sample was no option. While the evaluation of their third nudge is based on the full sample, the first and partly the second nudge could be only evaluated based on a limited sample. The lower power of the first two nudges limits the robustness of the analysis.

Second, we observed a relatively low participation rate of some households in the surveys and in the application that conveyed the nudges. For the lack of participation in the surveys, we encourage the missing participants to take part by sending them reminders. We also considered incentives for participation in the form of small gifts. For the lack of interaction in the apps, we considered the level of interaction in the analysis when the kind of nudge and the availability of app data enabled it.

Third, state-of-the-art information and communication technology enables the collection of high-resolution sensor data. Still, problems with the data transferred occurred between the pilot partners' households and the decentral data platforms and between the decentral and central data platforms based on Grafana. For instance, during some weeks in August 2022, no sensor data were transferred to the central data platforms. Thanks to the commitment of domX, this issue was resolved in a timely manner, and parts of the data were restored.

Fourth, measurement errors can create outliers that distort the analysis. We approached this issue with energy and statistical knowledge. On the one hand, illogical measurements, such as negative values, generation values beyond the installed photovoltaic capacity, or consumption values beyond the installed household capacity, were excluded. On the other hand, common outlier identification methods, such as those based on the interquartile range (IQR), were applied.

Despite the described mitigation measures, data gaps can occur. We considered three steps to handle those. First, aggregating collected data (e.g., daily values) limits the impact of single periods with missing data. We decided to use mean values instead of summed values since the number of missing observations distorts the mean less than the sum. Second, we considered interpolating data gaps by forward interpolation, backward interpolation, and nearest neighbor search. Since interpolations distort the measured data, we decided to handle the remaining data gaps after the first step with the third step: accepting an unbalanced data set.

3 German pilot: Increasing self-consumption

In the following section about the German pilot, three pilot-specific hypotheses are tested based on sensor data. The two general hypotheses on intention and motivation are tested based on survey data. As outlined in Table 3-1, we are able to confirm the three pilot-specific and one general hypothesis. The section references in the table guide the reader to the analysis based on which the hypothesis is tested. Additionally, we provide changes in the effect over time in an event study.

Table 3-1: Tabular summary for the German pilot

Hypotheses		Based on	Outcome	Section reference
DE1	Nudges are effective in increasing the self-consumption of participants.	DiD with sensor data (daily aggregation, baseline vs. nudge)	✓	3.2.2Nudge effects
DE2	Nudges are more effective in increasing the self-consumption of participants with controllable electric vehicles than of the ones without.	DiD with sensor data (peak – off-peak aggregation, baseline vs. nudge)	✓	3.2.3, 3.2.5
DE3	Nudges are effective in reducing the overall electricity consumption of participants.	DiD with sensor data (daily aggregation)	✓	3.2.2
All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	✓	3.3.1
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	–	3.3.1

3.1 Pilot-specific research design

3.1.1 Hypotheses

The 111 participants in the German pilot are prosumers with their own rooftop photovoltaic plants. The nudges were designed to encourage the participants to use their self-generated electricity more extensively and more effectively:

DE1: Nudges are effective in increasing the self-consumption of participants.

More effective usage can be realized by shifting their consumption or increasing it during hours of self-generated electricity. A web portal with information on household consumption and generation supports them to do so. Additionally, 39 out of the 111 participants own an electric vehicle (EV) that can be controlled by a smart charging app. The smart charging app aims to increase the EV's usage of self-generated electricity, considering the charging settings given by the household (e.g., departure time, targeted state of charge). Given the additional means to increase the self-consumption, we introduce another hypothesis:

DE2: Nudges are more effective in increasing the self-consumption of participants with controllable electric vehicles than of the ones without.

In the following, we call the sub-group with controllable EVs EV group and the one without PV group.

The enhanced knowledge and awareness about the households' electricity consumption also aims to decrease their overall consumption of electricity:

DE3: Nudges are effective in reducing the overall electricity consumption of participants.

3.1.2 Analytical strategy

For the analysis of the German pilot, we employ a DiD estimation strategy with two-way fixed effects. This strategy is well established in economics to capture the causal effects of interventions with panel data. The regression is set up to account for differences across groups ("first difference") by adding household fixed-effects, i.e., allowing for a separate intercept term for each household. In practical terms, this cancels out factors that are common for a household over time, such as the size of the PV plant, their energy-consumption habits, and the energetic properties of the building. Then, a second set of fixed-effects is added for each time period (day). These day-specific intercepts then cancel out factors that are common to all households on a given day, the main factors being the weather and the developments during the energy crisis. The assumption here is that the news and policy changes are common to all households, so the time-fixed effects can absorb their effect. This is the "second difference". The DiD then captures only what is left over: the treatment effect. This treatment effect measures the differential development, i.e., whether the treatment group changed more (or less) than the control group between the baseline and the intervention period.

The DiD method reliably estimates the causal effect of nudging as far as the parallel trends assumption is respected: the daily electricity consumption levels of the two groups have to follow the same time trend during the baseline period, although they do not have to be the same in absolute terms. Hence, we start with a descriptive analysis of the data to understand the sample composition and verify that the data are indeed suitable for the chosen analytical strategy. This is critical because each nudge is compared to the baseline period as the reference period, not the previous nudge period. The treatment effects always compare the development between the baseline and the respective nudge.

The hypotheses target two outcome variables: household consumption and self-consumption. Household consumption is the energy a household consumes excluding energy used for charging electric vehicles. Self-consumption is that part of the consumed energy that comes from the household's own production. In the section Nudge effects, we conduct the DiD estimation for both outcome variables to test hypotheses 4.1 and 4.3. Additionally, we conduct a special analysis for nudge 3, the opt-in nudge, which is a special kind of default nudge that requires an initial confirmation to be activated. This nudge targeted to EV owners is an opt-in feature: once households activate it, the opt-in is smart charging instead of regular charging. Hence, this nudge requires low interaction, but it is only effective for users who actually change the opt-in settings to activate the smart charging feature. In contrast to the other two nudges, the response of the household to the nudge in the app can be clearly linked to a change in consumption behavior. For instance, while monitoring additional information of one nudge can lead to no change, an immediate or a delayed change of behavior, the activation of the automated opt-in is immediately recognizable in the consumption behavior. The clear link between the action in the app and the consumption behavior encouraged us to dive more deeply into the mechanism behind this third nudge.

To test hypothesis 4.2 about the specific effect of households owning an electric vehicle, we re-estimate the same DiD model described above but distinguish between the EV and the PV groups to assess whether these sub-groups respond differently to the nudges.

Acknowledging the challenge of fatigue effects (decreasing interest of nudged subjects over time), which is reported in the literature[38], we test whether the nudge effects develop throughout the intervention period. This is done with an event study design, which essentially estimates a separate treatment effect for each day during the intervention. This allows an analysis of whether the nudging effect is constant throughout the intervention or whether there are time trends within the intervention. A caveat is that this estimation strategy is very demanding because the cross-section (number of households) is much smaller than the time series (number of days). We anticipate that there may not be enough power to get very precise estimates for each individual day, but the strategy should still reveal whether there are broad patterns over the intervention period.

3.1.3 KPI & data

We calculated two metrics to assess how participants responded to the nudging interventions: one that provides an absolute assessment (self-consumption) and another that offers a relative evaluation of consumption compared to the overall consumption (autarky rate). It's worth noting that, while in previous studies, the latter metric has been referred to as self-sufficiency-rate, we've chosen to label it as autarky-rate to prevent any potential terminology confusion with research focused on sufficiency.

The absolute measure takes into account both responses to the nudging intervention, increasing and shifting of consumption. At the same time, it may be influenced by random changes in consumption, such as vacations or construction activities. The relative measure factors in these consumption fluctuations, even absorbing any additional consumption prompted by the nudging interventions. For self-consumption, we log-transform the outcome variable to mitigate the impact of outliers.

3.1.4 Nudging interventions

Two tools, a web portal and a smart charging app were utilized to present the nudges to the participants. Notably, the smart charging app was exclusively accessible to participants with controllable EVs.

We consider learning and fatigue effects when determining the sequence of the three nudging interventions. We establish previously implemented nudges as new basic settings to handle learning effects during the three nudging periods. Subsequently, we compute the incremental change for each new nudge (e.g., the difference in consumption between nudge 1 and 2). To prevent fatigue effects, we start with the nudge that demands the most interaction ("feedback") and decrease this demand over the nudging periods (see

Figure 3-1).

[38] Jared J. Cash, Alert fatigue, American Journal of Health-System Pharmacy, Volume 66, Issue 23, 1 December 2009, Pages 2098–2101, https://doi.org/10.2146/ajhp090181

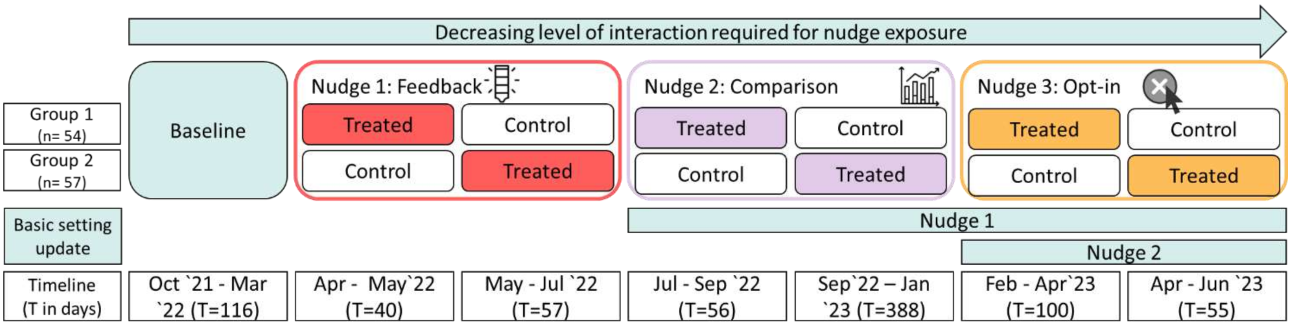


Figure 3-1: Experiment outline for the German pilot

Both tools introduced a new dashboard for the initial nudge (Figure 3-2), which provided feedback. This dashboard employed straightforward indicators coupled with color-coded signals to categorize the participant's current self-consumption behavior as either favorable (green) or unfavorable (red).

In the case of the second nudge (Figure 3-2), involving comparisons, participants' present self-consumption levels were compared to their previous ones through a bar chart. To promote an increase in self-consumption, forecasts for self-generated electricity and consumption recommendations were presented.

As the third nudge (Figure 3-2), an opt-in nudge, a new charging mode was introduced for participants with controllable EVs. The existing charging mode had been designed to schedule the charging process to maximize self-consumption based on specific target state of charge and departure time parameters. The new charging mode, however, would only charge the EV using self-generated electricity unless overridden by setting a target state of charge and departure time. Once participants accepted this new charging mode in the tool, it would be automatically activated whenever the EV was plugged in at home.

Simultaneously, an additional nudge was introduced for all participants, even those without controllable EVs. This additional nudge re-used the information from the two previous nudges and presented it in the form of a downloadable energy report, thus ensuring engagement for all participants.

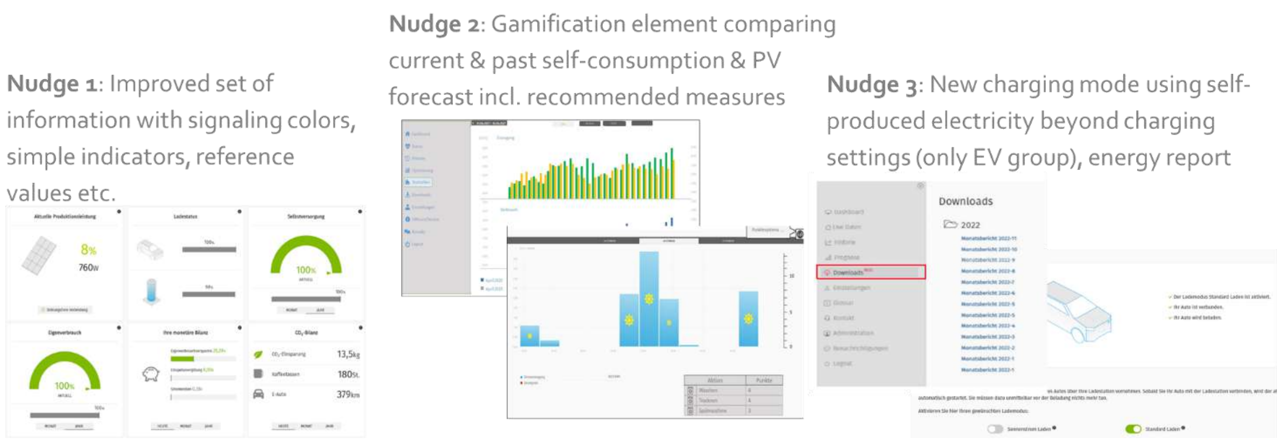


Figure 3-2: Nudges of the German pilot

3.2 Testing pilot-specific hypotheses

3.2.1 Descriptive statistics

For the analysis, the sensor data are aggregated at a daily level. The key condition for the research design is that the treatment and control groups are roughly comparable with respect to their autarky rate. The descriptive analysis supports that the two groups are indeed similar in their outcomes, so the randomized assignment to groups was successful. In Figure 3-3, we plot the time-averages of the two KPIs, the autarky rate (range: 0 to 1) and self-consumption (in Wh, reported as the mean hourly value over a 24-hour period). Clearly, the two groups have similar patterns. Most importantly, the two groups do not differ during the baseline measurements. Notably, the variation of the autarky rate with time is highly weather-dependent, as there is a common up-and-down pattern that matches the pattern of solar radiation over each day. Throughout the year, there is a clear seasonal pattern, with higher averages in the summer months. Comparing the two KPIs, it appears that they follow the same trend, but the autarky rate is less volatile than self-consumption. This was expected and made autarky rate the preferred outcome for the regression analysis. The summary statistics in Table 3-2 support the insights from the visualization: both groups are comparable, have sufficient observations to allow regression analysis, and the values fall within the expected range. Nevertheless, the statistics also show that the prosumers in our sample consume more energy than the average German household, and the high standard deviation (SD in column 3) indicates substantial heterogeneity across households.

The particularity of the German pilot is the split into prosumers with controllable EVs and those without. Figure 3-4 shows this division into the sub-groups: the average for all households (left), only the EV group (middle), and the PV group (right). The data are aggregated at a weekly level for better visibility. The data are very similar across the sub-groups. There is no indication that the EV group behaves differently from the PV group from the onset. An important factor here is that e-mobility is high in the sample in general. Even within the PV group, many households still own an electric vehicle, which explains that there is not strong discrepancy between the sub-groups.

Table 3-2: Summary Statistics by Group

	Mean	SD	Min	Max	Obs
Group 1 (n = 54)					
Consumption [Wh]	755.58	586.12	0.05	7503.91	23029
Self-consumption [Wh]	445.49	358.8	0	3863.78	23029
Autarky rate [percentage]	0.55	0.24	0	1	23029
Household Consumption [Wh]	646.98	433.05	0	8545.54	22245
Group 2 (n = 57)					
Consumption [Wh]	720.18	565.57	0	5987.11	24010
Self-consumption [Wh]	459.33	370.57	0	4188.53	24010
Autarky rate [percentage]	0.60	0.23	0	1	24010

Household Consumption [Wh]	607.08	466.46	0	5987.14	23032

Notes: Descriptive statistics for estimation sample from January 2022 to June 2023 at daily aggregation. Self-consumption is the difference between total consumption and output to grid. Autarky rate is the ratio of self-consumption to total consumption. Household consumption is energy consumed without EV charging.

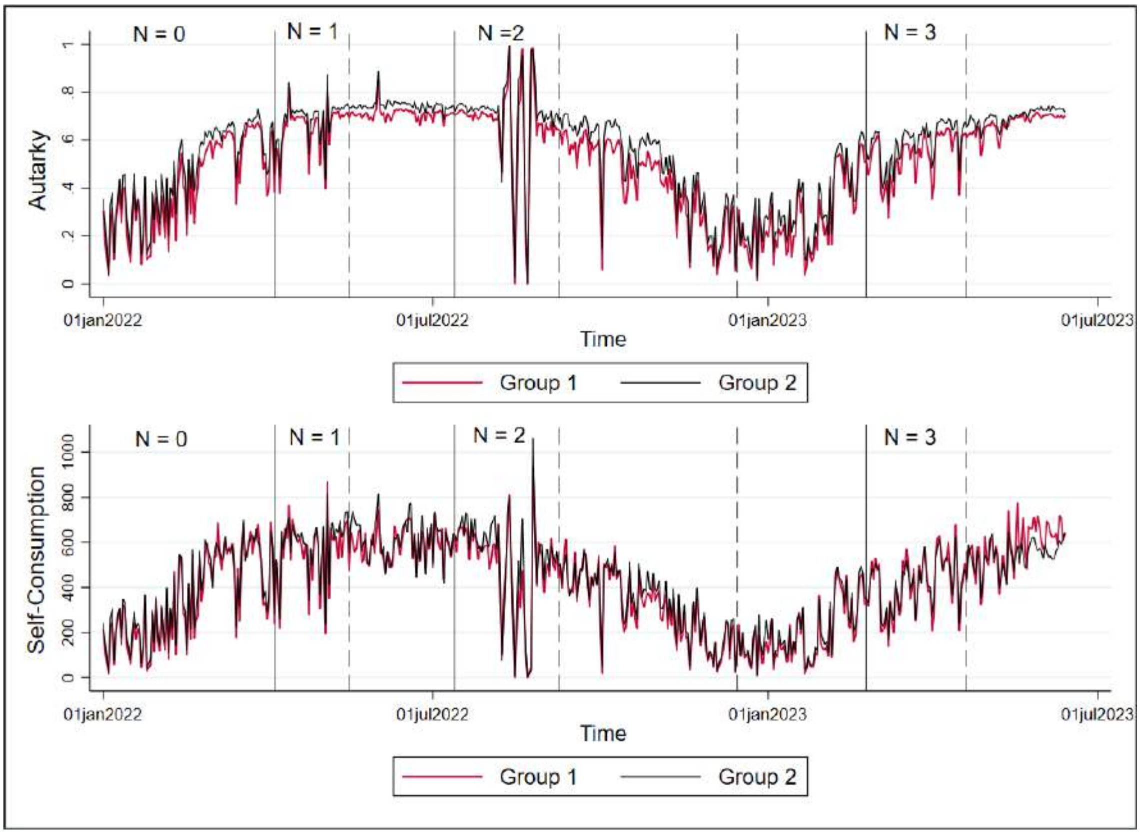


Figure 3-3: Indicators of self-consumption by group

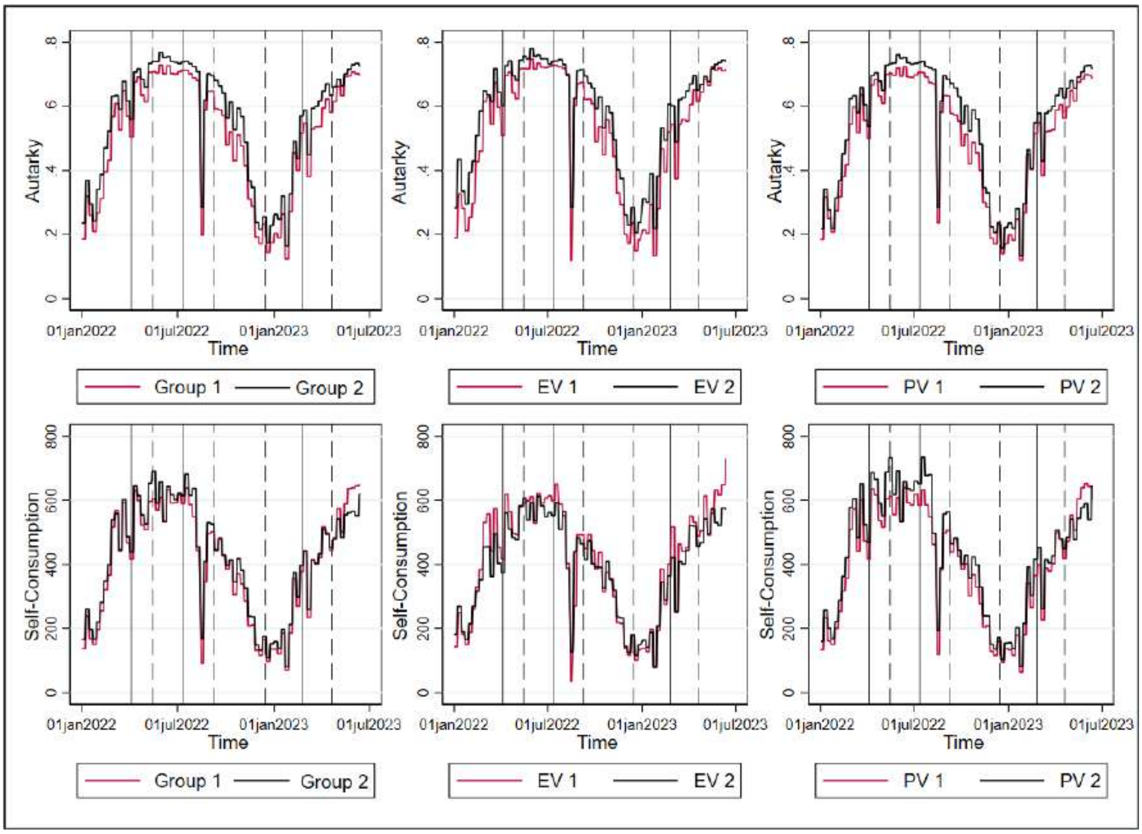


Figure 3-4: Indicators of self-consumption in EV and PV sub-groups

Besides the two self-consumption indicators depicted above, the analysis considers household consumption as a third outcome variable. Figure 3-5 plots the patterns for both groups over time. Household consumption is in the upper panel. For comparison, total energy consumption is included in the lower panel. The two groups follow similar trends over time, with less volatility in household consumption than in total consumption, as expected. Notably, there is a slight divergence during the second phase of nudge 2, where group 1 records higher household consumption than group 1. This appears to be seasonal, as the groups converge again during spring.

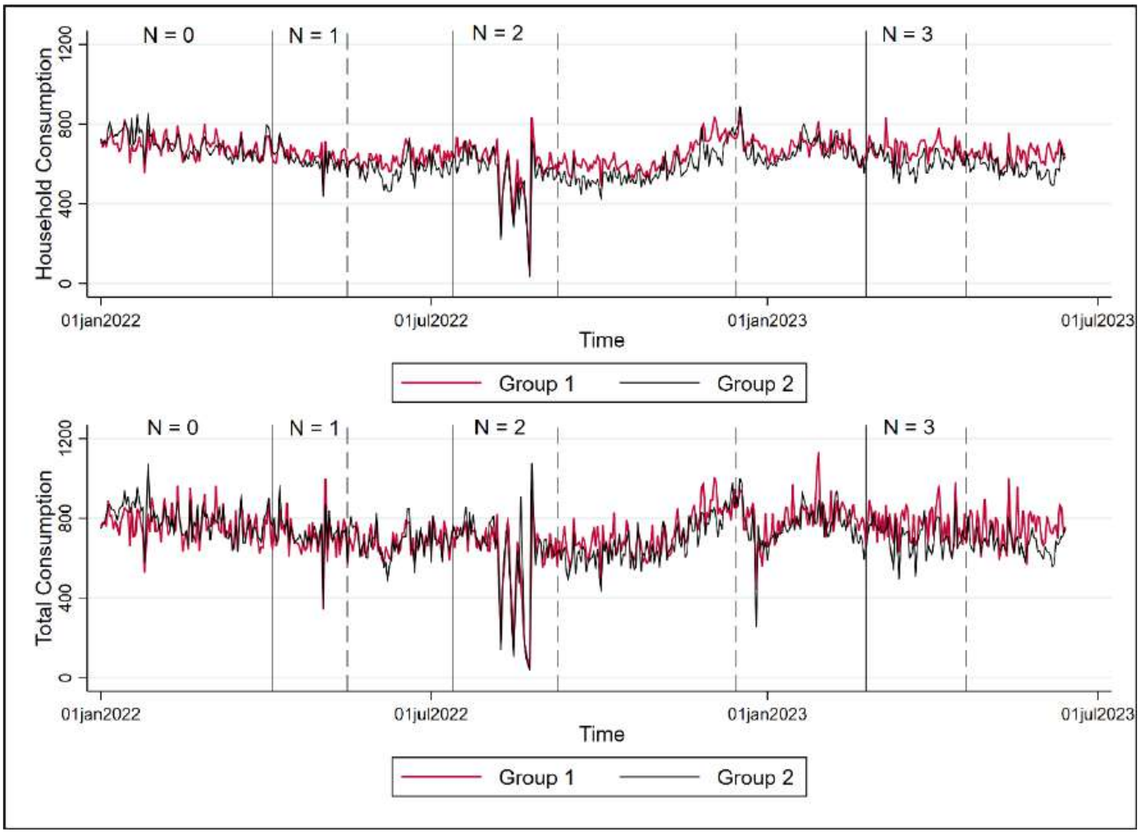


Figure 3-5: Indicators of energy consumption by group

3.2.2 Nudge effects

Table 3-3 reports the main results for nudges 1-3. It is structured according to the two main research questions: effectiveness of nudging interventions and differences across the three nudges. The left part of the table shows treatment effects that indicate how effective the nudges are for optimizing self-consumption and reducing household consumption. These are the main results obtained with the DiD strategy described above that evaluates the nudging periods against the baseline. The right part of the table compares the three nudges to answer the question of whether some nudges work better than others. In horizontal order, Panel A reports results for group 1, which received the treatment first.

The first two nudges deliver similar results and are therefore discussed together. For nudge 1 (feedback) and nudge 2 (comparison), we find small, positive treatment effects regarding the two self-consumption indicators. The reported coefficients for autarky indicate that the nudging treatments increased autarky by ca. two percentage points. Since autarky is measured on a scale from zero to one, it is useful to put this estimate in relation to the mean autarky rate in the sample, which is an autarky rate of 0.55. At this level, the 2-point improvement corresponds to an increase of 3.8% on average. Regarding self-consumption, the coefficients can be interpreted directly as changes in percent due to the logarithmic transformation. Self-consumption increased by 2.9% due to the feedback nudge and by 2.8% due to the comparison nudge.

However, the estimates for self-consumption are much less precise than the estimates for autarky. While there is strong statistical significance for autarky, the results for self-consumption are not statistically significant at the conventional 5%-level of confidence. We attribute this to the higher volatility in the time-series of self-consumption that emerged from the descriptive analysis.

For household consumption in column 3, both nudges 1 and 2 led to a significant reduction as expected. Nudge 1 led to a reduction by 3.9% on average, nudge 2 to a reduction by 5.2% (again, household consumption was log-transformed). Taken together, the results indicate that households managed to increase self-consumption and simultaneously decrease household consumption with the nudging intervention. This is encouraging because it was not clear ex ante whether both would be possible at the same time: rebound effects would predict that higher self-consumption is channelled to additional household consumption.

Finally, all tests for a difference between the effectiveness of nudge 1 and nudge 2 are insignificant (in technical terms: Wald tests fail to reject the null hypothesis of no difference). This is indicated by the very small test statistics and the corresponding high p-values in the right part of the table. Hence, we find no evidence that the feedback nudge is more effective than the comparison nudge or vice versa.

The results differ for nudge 3. Here, the coefficient for autarky is very close to zero: nudge 3 appears to have no economically relevant effect on the autarky rate. By contrast, the nudge has a strong positive effect on self-consumption, which increases by 11.1%. To understand these effects, it is useful to recall that autarky is the ratio of self-consumption to total energy consumption. If autarky is unaffected but self-consumption increases, the intuitive explanation is a concurrent increase in total consumption. We tested this hypothesis and indeed found evidence for such an increase in *total* energy consumption. Total energy consumption is energy used in the household plus energy used for charging the EV. By contrast, there is no significant effect on household consumption alone, as shown in column 3. Given that nudge 3 has the added functionality for smart charging for users with controllable EVs, we interpret this result as suggestive of additional charging during nudge 3. Whether this constitutes a rebound effect depends on whether households shifted charging to the home or increased charging overall. Not surprisingly, the tests in the right panel support the view that nudge 3 worked differently than nudges 1 and 2 (all p-values <0.001 for the two self-consumption indicators and <0.05 for household consumption).

Panel B in the middle reports the corresponding results for the second group. Note that the DiD estimator compares the relative development across groups. Hence, when group 2 is treated, the control group is group 1, which had previously seen the respective nudge, but is no longer exposed to it. This changes the interpretation of the coefficients if nudges have long-term effects. For nudge 1 (feedback), the treatment effects are unexpectedly negative. Autarky decreased by 3 points, and self-consumption decreased by 6.8%. Hence, the order of magnitude is rather similar to group 1, but the direction is reversed. For nudge 2 (comparison), the treatment effects are positive. The estimated increase in autarky by 1.4 points corresponds to a 2.3% increase relative to the sample mean (0.60). This result is only slightly smaller than for group 1, indicating that the treatment worked similarly for both groups. This holds when considering self-consumption as an alternative outcome. As with group 1, the precision of the estimates is weaker for self-consumption. For household consumption, the nudge 1 coefficient is positive but not significantly different from zero. For nudge 2, the effect is unexpectedly negative, indicating a 2.8% decrease for group 2 relative to group 1. By contrast, the treatment effects for nudge 3 are negative for both self-consumption indicators. Autarky decreases by 2.4 points (or 4% relative to sample mean). Self-consumption decreases by 15.7%, which is a substantial drop that does not conform to expectations. The result for nudge 3 is close to zero and insignificant, which aligns with group 1. The formal tests for the pairwise comparison of the three nudges all reject the null hypothesis of equality, with the exception of nudges 2 and 3 for household consumption. This mean is that the three nudges all exert different effects. These tests essentially confirm the picture drawn out of the coefficients: for group 2, we cannot provide strong comparative results across the three interventions because the effect sizes are inconsistent in both magnitude and sign.

We conducted a number of robustness checks to rule out that the unexpected results are driven by outliers or other changes from individual households but found that the sign of the effect remains stable. Therefore, the most likely explanation is the order of treatment in the design. For discussion, the measurements of the second group are impacted by long-term effects for the first group. For example, if the first group adopted new habits regarding their energy-saving behavior, this would impact the results. The treatment effect says whether the treated group changed more (or less) than the other group, not whether they changed at all. Among economists, there is a recent and ongoing debate regarding the interpretation of "staggered treatments", where groups receive the same treatment at different times. While the experiment was already underway, new insights from statistics were published that show how sensitive statistical methods can be when there is no "clean" (i.e., never-before-treated) control group (see Section 2.4).

Therefore, we have more confidence in the results of the first group because these households can be compared to those that have never seen the treatment. Regardless, it would be interesting for future research to better understand such learning effects and why we only find them for some nudges, but not consistently for all interventions.

In Panel C at the bottom, we report the R^2 of the model as a measure of model fit. R^2 is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. The values that the model explains ca. 78% of the variation in autarky, and ca. 70% of the variation in self-consumption. These values are highly satisfactory against the established literature. The R^2 also supports our choice of the two-way fixed effects model. If we estimate the model without any fixed-effects, the R^2 is much smaller. This means that a large fraction of the variation in households self-consumption patterns is driven by fluctuations over time and across households that are not related to the nudging. Across households, the fixed effects would absorb persisting differences in energy-saving behavior, as well as differences in PV size, other equipment and household size that determine energy usage. Over time, one might expect that weather is the main driver.

However, political developments, energy prices and other time-dependent factors might also be a factor, especially considering the turbulence brought about by the energy crisis in 2022. We explore this more-than-weather hypothesis by replacing the time-fixed effects with weather variables (radiation and temperature). With this substitution, the R^2 still drops by more than 15 percentage points relative to the fixed-effects model. This indicates that prosumers energy usage is not a pre-determined function of weather, but that the households in the sample also react to other time-dependent developments in their environment. The R^2 is lower for household consumption than the two self-consumption indicators, which suggests that there is more random fluctuation for this variable. This could arise because self-consumption is more pre-determined by the weather, which is correctly accounted for by the time-fixed effects, whereas there is consumption patterns in the household are noisier as a result of everyday decision-making.

Table 3-3: Full list of coefficients and p-values for nudge comparison for the German pilot

Coefficients for Treatment Effects (Std. Errors below)				F-Statistics from Wald Test (P-values below)			
	(1)	(2)	(3)		(4)	(5)	(6)
	Autarky rate	Self-Consump.	Household Consump.		Autarky rate.	Self-Consump.	Household Consump.
Panel A: Results for Group 1							
Nudge 1	0.0209***	0.0291*	-0.0385***	*Nudges 1 and 2*	0.00	0.00	0.78
	(5.57)	(1.73)	(-2.79)		(0.955)	(0.961)	(0.376)
Nudge 2	0.0212***	0.0280	-0.0524***	*Nudges 1 and 3*	52.27	20.41	7.01
	(5.05)	(1.32)	(-3.36)		(0.000)	(0.000)	(0.008)
Nudge 3	-0.00935**	0.111***	-0.00859	*Nudges 2 and 3*	44.33	13.09	3.97
	(-2.26)	(6.64)	(-0.59)		(0.000)	(0.001)	(0.046)
Panel B: Results for Group 2							
Nudge 1	-0.0299***	-0.0684***	0.0101	*Nudges 1 and 2*	196.92	40.23	2.73
	(-9.32)	(-4.27)	(0.76)		(0.000)	(0.000)	(0.098)
Nudge 2	0.0140***	0.0264*	0.0281**	*Nudges 1 and 3*	3.27	25.41	9.50
	(3.87)	(1.84)	(2.47)		(0.071)	(0.000)	(0.002)
Nudge 3	-0.0240***	-0.157***	-0.00612	*Nudges 2 and 3*	110.05	133.30	1.55
	(-6.45)	(-9.15)	(-0.46)		(0.000)	(0.000)	(0.214)
Constant	0.574***	5.815***	6.250***				
	(792.08)	(1678.54)	(2435.36)				

Panel C: Regression Model Specification							
R^2	0.778	0.702	0.655				
R^2 Without FE	0.255	0.111	0.021				
R^2 Weather	0.639	0.525	0.592				
N	46409	45928	44671				

Notes: *Results on left side are difference-in-differences estimation for dependent variables autarky (0 to 1 ratio), self-consumption, and household consumption (log-transformed). Household consumption is total consumption excluding EV charging. Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Right side shows Wald test, with p-values in parentheses below to compare effects across nudges. The alternative R2 measures in Panel C refer to specifications that include no fixed effects ("Without FE"), and a specification with household fixed-effects and variable weather controls ("Weather"),*

3.2.3 Deep dive: Active vs. passive participants in nudge 3

Considering that nudge 3 is a special case with particular relevance for the EV group, we conduct additional analysis to better understand how prosumers use this feature. Overall, the analysis of the opt-in nudge provides a two-part result. On the one hand, the opt-in nudge was highly effective in increasing self-consumption during the spring. We find stable positive effects of around 15% for the first group among those users who activated the feature. On the other hand, during summer, the opt-in nudge appears to be less effective, and we can attribute this to the relative abundance of solar power available to households. Further research work is needed to understand how saturation affects household behavior, as our results are power-constrained and can only provide an indication. Across both groups, the uptake of the feature was unexpectedly low and staggered throughout the intervention period.

In the first step, we check which users actually activated the smart charging function. Surprisingly, many consumers did not take advantage of the offer at all or activated it rather late in the intervention phase. In EV group 1, only 9 out of 18 households in EV group 1 actively used the feature. In EV group 2, all 18 households eventually activated the feature, but only one household activated it within the first week, and 9 out of 18 households only activated it within the last two weeks.

From this information, we categorize active households as those that had activated the feature on a given day. We then investigate whether consumption patterns shift throughout the day when consumers activate the opt-ins. If consumers who activate the feature move to longer plug-in times, the smart charging feature would primarily use energy at times of high PV generation. We would accordingly expect higher self-consumption during the midday peak of solar radiation by those who have activated the smart charging feature. This can be tested formally with a regression that uses the behavior of inactive participants as a control group. We run this model with all households and then again with only the inactive EV sub-group as a comparison. The latter ensures that we do not pick up general differences between the EV and PV participants, which would bias the estimates.

The data are compiled at hourly frequency and the day is divided into three blocks: AM (6:10:00), Midday (11:15:00) and PM (16:20:00); the night hours without PV generation are dropped. Because the main results suggest that nudge 3 also affected total energy consumption, we add this outcome to the previous KPIs of autarky and self-consumption.

The results for group 1 are shown in Table 3-4 below. The coefficients indicate whether the active group shifts more than the control group (formally, this is estimated with an interaction term in the regression model). The results do not support an effect on autarky – the coefficients are close to zero and not statistically different from zero (columns 1 and 2). However, the active participants have substantially higher self-consumption during the midday peak (columns 3 and 4). The coefficients indicate that self-consumption increases by ca. 16% when the smart charging feature is activated. Notably, there is a simultaneous increase in total consumption by a similar amount (columns 5 and 6). In the row below, we also explore what happens in the evening hours. Here, the coefficients are mostly negative, but insignificant throughout the entire row. There is no evidence that the opt-in nudge has any effect during the evening. The shift occurs only during the midday window. Overall, the results support the hypothesis that the opt-in nudge allowed users to substantially increase self-consumption after activating the smart charging. However, the caveat is that only a few users took advantage of this option in the first place.

Table 3-4: Intraday shifts for group 1

	(1)	(2)	(3)	(4)	(5)	(6)
	Autarky Rate		Self-Consumption		Total Consumption	
Active x Midday	**-0.00217**	**0.0132**	**0.165****	**0.157****	**0.165****	**0.135***
	(-0.15)	(0.94)	(2.28)	(2.17)	(2.26)	(1.84)
Active x PM	**0.0114**	**-0.00871**	**-0.0587**	**-0.130**	**-0.0117**	**-0.0313**
	(0.42)	(-0.32)	(-0.39)	(-0.87)	(-0.15)	(-0.39)
Midday	0.184***	0.173***	0.721***	0.802***	0.176***	0.260***
	(80.26)	(44.97)	(66.91)	(42.21)	(23.91)	(20.68)
PM	-0.0209***	0.00600	0.118***	0.260***	0.0972***	0.170***
	(-7.42)	(1.29)	(8.59)	(11.25)	(13.83)	(14.69)

R^2	0.352	0.322	0.200	0.193	0.250	0.196
Obs.	72802	26325	69347	25270	72802	26325
Control group	All	EV only	All	EV only	All	EV only

Notes: *Regression testing for intra-day shifts during Nudge 3 for group 1. Data at hourly frequency. Baselevel is AM (6:10:00). Active is an indicator for interaction with the app. Dependent variables autarky (0 to 1 ratio), self-consumption, and total consumption (log-transformed). Total consumption includes EV charging. Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels:* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The same analysis is then repeated for group 2, and the results are shown in Table 3-5. Broadly speaking, the results for group 2 are inconclusive, which provides further indication that there are important differences between the two groups. In summary, there are three unexpected differences relative to group 1. (I) the effects are distributed over the day rather than concentrated in the midday, (ii) the results are sensitive to the choice of a control group, (iii) there are stronger effects for total consumption than self-consumption.

We first present the results, and then discuss the context in comparison to group 1. For autarky in columns 1 and 2, there are small, negative effects during the midday peak in the range of 3 to 4 percentage points. By contrast, there are small, positive effects for the PM hours in the range of 1 to 3 percentage points. The active users appear to have slightly lowered self-consumption in the middle of the day, but then increased it during the evening. For self-consumption, we find that self-consumption increased by 4.85% during the midday peak, but by 12.2% during the PM hours (column 3). The PM effect is highly statistically significant, the midday peak is only significant at the 10% confidence level. When the control group is restricted to only the EV participants, the results are insignificant for both terms (column 4), which indicates that we do not find any differential development between the active and the non-active EV households in group 2. This is in sharp contrast with group 1, which showed a substantial increase for the active users that was concentrated in the midday window. For total consumption (column 5), there is a solid increase of 16.4% during midday, and a smaller, but still significant increase of 8.5% during the PM window. When only the EV participants are used as the control group (column 6), the pattern persists qualitatively, but the effect sizes decrease, and the PM effect is no longer significant.

Comparing the model fit, i.e., the R^2, in the bottom panel between Table 3-4 for group 1 and Table 3-5 for group 2, it appears that model fit is somewhat weaker for group 2 than for group 1. The drop in model fit is more pronounced when only the EV participants are used as the control group (columns, 2,4 and 6, respectively).

Table 3-5: Intraday shifts for group 2

	(1)	(2)	(3)	(4)	(5)	(6)
	Autarky Rate		Self-Consumption		Total Consumption	
Active x Midday	-0.0430***	-0.0314***	0.0485*	-0.0167	0.164***	0.0890***
	(-10.93)	(-6.99)	(1.73)	(-0.53)	(5.88)	(2.89)
Active x PM	0.0273***	0.0101*	0.122***	0.0176	0.0845***	0.00634
	(5.12)	(1.68)	(5.32)	(0.66)	(3.78)	(0.25)
Midday	0.0250***	0.0210***	0.0593***	0.162***	0.0183**	0.114***
	(17.05)	(8.00)	(6.51)	(9.32)	(2.02)	(6.57)
PM	-0.105***	-0.0793***	-0.433***	-0.291***	-0.249***	-0.151***
	(-50.68)	(-21.97)	(-43.22)	(-15.72)	(-30.57)	(-10.00)
R^2	0.295	0.258	0.229	0.163	0.221	0.122
Obs.	72667	26439	70793	25990	72667	26439
Control group	All	EV only	All	EV only	All	EV only

Notes: *Regression testing for intra-day shifts during Nudge 3 for group 2. Data at hourly frequency. Baselevel is AM (6:10:00). Active is an indicator for interaction with the app. Dependent variables autarky (0 to 1 ratio), self-consumption, and total consumption (log-transformed). Total consumption includes EV charging. Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels:* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To interpret the findings, the key difference between the groups is the timing of the intervention. Group 1 received the nudge in the spring (24 February – 19 April), whereas group 2 was treated in the early summer (20 April, 13 June). Taking the middle of the intervention period as a reference value for group 1, the gap from sunrise to sunset is about 12 hours, with sunset coming well before the end of the PM window (March 22: sunrise 06:24, sunset 18:40, day length 12:16). For group 2, the day lasts 15 and a half hours, and sunrise comes after the end of the PM window (May 17: sunrise 05:34, sunset 21:09, day length 15:35).

The timing has two implications. First, there is more opportunity for self-consumption during the PM window. Hence, the effects of the smart charging feature can be spread throughout the day rather than concentrated in the midday peak. This explains why we find effects for the PM window only for group 2.

Assuming the total number of charging procedures does not change substantially, this diverts from the midday concentration we recorded for group 1. Second, group 2 also has higher solar radiation during their treatment. Group 2 had a mean daily autarky of 0.55 during their treatment, whereas Group 2 had a daily mean of 0.71. During sunshine hours, the values approach 1 for many households in the sample. We explored the time series for all three outcomes individually for each household in EV group 2 and indeed found very high autarky rates during nudge 3 for a large fraction of the treatment group. Incidentally, the EV2 group has larger PV plants (mean of 8.1 vs 7.37 in EV 1), and larger batteries (mean of 7.28 vs. 6.32 in EV 1). Battery and PV size are key determinants of the autarky rate from a technical perspective.

From the methodological perspective, this is a major problem for estimation. When households are already close to the upper bound of autarky without the intervention, there is little room for further improvement. The maximum gains in additional self-consumption are relatively small, so a large sample would be needed to credibly estimate the effect. This is a basic problem with statistical power in regression analysis, which is especially pronounced here because many households activate the feature late, so there are few observations with activity in the overall sample. Adding in the fact that optimization is spread over more hours, the total effect is divided between the two windows, which makes it even more difficult to obtain enough power for significant estimates. Against this background, it makes sense that the results with the smaller control group (columns 2,4,6) tend to be insignificant and out of line with those obtained from the larger sample (columns 1,3,5). In practical terms, we interpret the findings as evidence of a rebound effect: households with the smart charging feature that were already close to full autarky, so it appears the response was at least partially diverted to an increase in overall consumption.

Finally, there is a possibility of learning effects, although we expect this to be less influential for nudge 3. Group 1 did not have access to the smart charging feature anymore, and it is not technically feasible to do this optimization "by hand". However, it is still possible that users saw in their energy report (also part of nudge 3) that EV charging has a strong impact, and thus became generally more strategic about their charging. With the limited sample and the seasonal trend, we are unable to estimate learning effects credibly in the context of the experiment, but we expect learning to be of minor importance in nudge 3 due to the low-interaction design of the intervention.

3.2.4 Time-dependent effects: Event study

The main results presented in Table 3-3 give the average effect over the entire treatment period. In an extension, we also explore whether the nudge effect changes over the study period. On the one hand, the technical setup required the participants to update the app before the new feature was displayed. If consumers don't update instantly, there would be a delay before nudging effects appear. On the other hand, the existing literature suggests that consumer engagement fades after an initial excitement. We do see in the app data, that most activity happens around the time of announcement and then levels off. In this case, the nudge effect would decrease over time. Figure 3-6 and Figure 3-7 show what happens in the first 20 days of around the introduction of each nudge. The event study design that is used for this analysis sets the day of implementation to 0 on the x-axis, so days before implementation are counted as negative (lag terms), and days after implementation are numbered in consecutive positive numbers (lead terms). The black circle gives the coefficient for each day with effect sizes on the y-axis, and the capped line reports the corresponding 95%-confidence interval. A coefficient is significantly different from zero when the confidence interval does not touch the red horizontal line. All measurements are relative to the control group as in the main result, so the graph shows for each day whether the treated group reacted more or less than the control group.

For group 1 in Figure 3-6, we do not find strong evidence for a distinct time pattern. The coefficients for nudges 1 and 2, are positive but largely insignificant. There are no clear patterns of upward or downward trends. For nudge 3, the pattern is mostly similar, although there are several reversals and generally less stable results. However, the previous results have shown that consumers to activate the feature at very different times and that some of the effect comes from intra-day changes, so the lack of a common trend is at least partly due to micro-patterns that the event study cannot capture well.

Hence, the results do not support either fatigue or delay in the nudge. Instead, the conclusion from the event study is that the nudge works continuously over the intervention period. As expected, the series for autarky is much less volatile than for self-consumption (note the different y-scales). In fact, it appears that the autarky series is more stable during the intervention than before for nudges 1 and 2, which indicates that behavior became more regular during the intervention. The lack of significance for individual days is attributed to the small sample size: there is less information available in a single day than over the entire period, so the model is less precise and gives larger confidence intervals. This is normal for a pilot of this size, but the coefficient series can still provide an indication of what is happening over time. For group 2 in Figure 3-7, the overall pattern is the same, which in this case assuages concerns that single-day anomalies drive the unexpected treatment effects from the main result. For nudge 1, the series is mainly negative but, again, very stable for autarky. For nudge 2, there are a few days with significant effects on autarky, which occur in the second week of the treatment period. For self-consumption, there is no clear pattern for either of the two nudges. For nudge 3, the outcomes are reversed: no clear pattern for autarky, a mainly negative series for self-consumption with scattered individual days that are significantly different from zero. Overall, the results match for both groups, as we find no evidence for common patterns in the time series that would match with fatigue effects or delayed onset. The small, average effects estimated for the main result credibly capture the underlying intervention period.

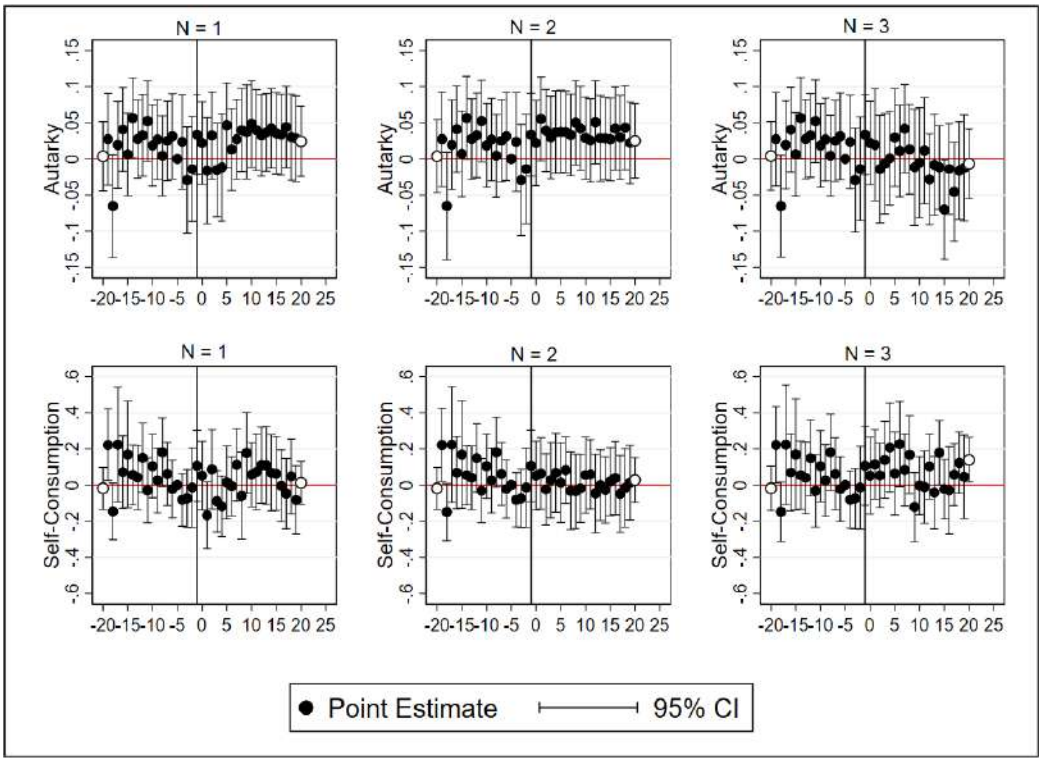


Figure 3-6: Event study results for self-consumption in group 1

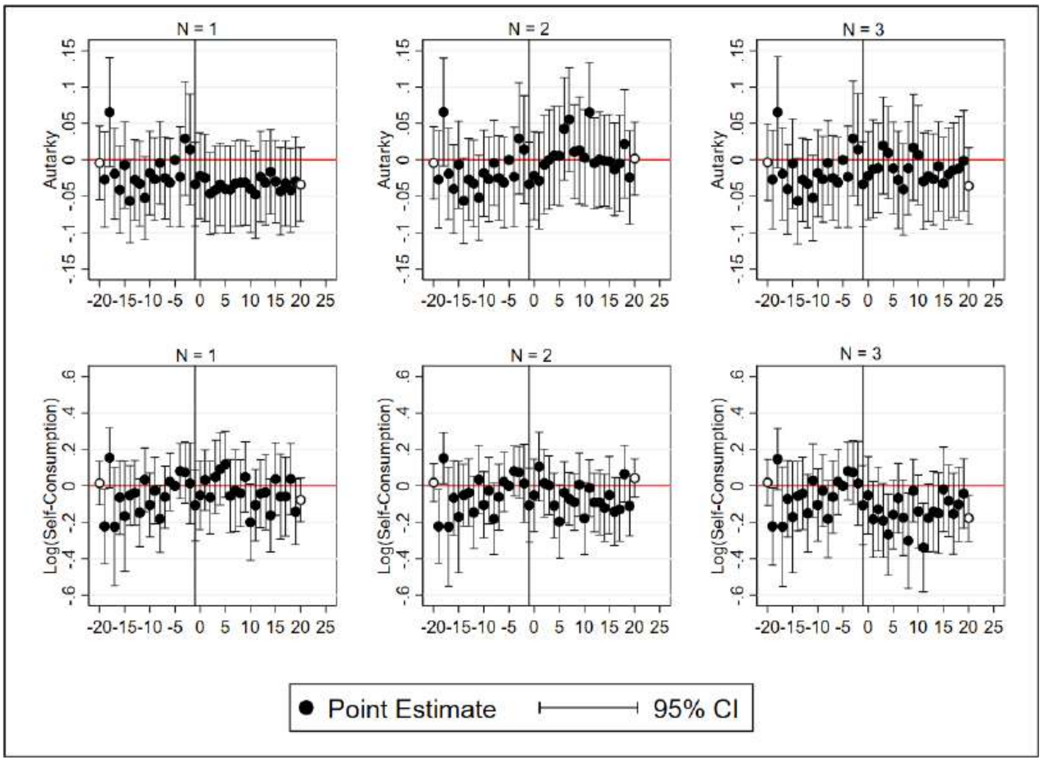


Figure 3-7: Event study results for self-consumption in group 2

The results are similar for household consumption, which is shown in Figure 3-8 below. Here, the upper panel reports results for group 1, the lower panel refers to group 2. The coefficients are again clustered around zero with no discernible patterns that would support fatigue effects. The confidence intervals are similar in size between self-consumption and total consumption, except for nudge 3 for group 2 (bottom right), where household consumption has wider confidence intervals, indicating more volatility, especially in the beginning. There is no clear explanation for this in the raw data, but several households reported changes to their household equipment in the last survey wave, which would be consistent with the observed pattern.

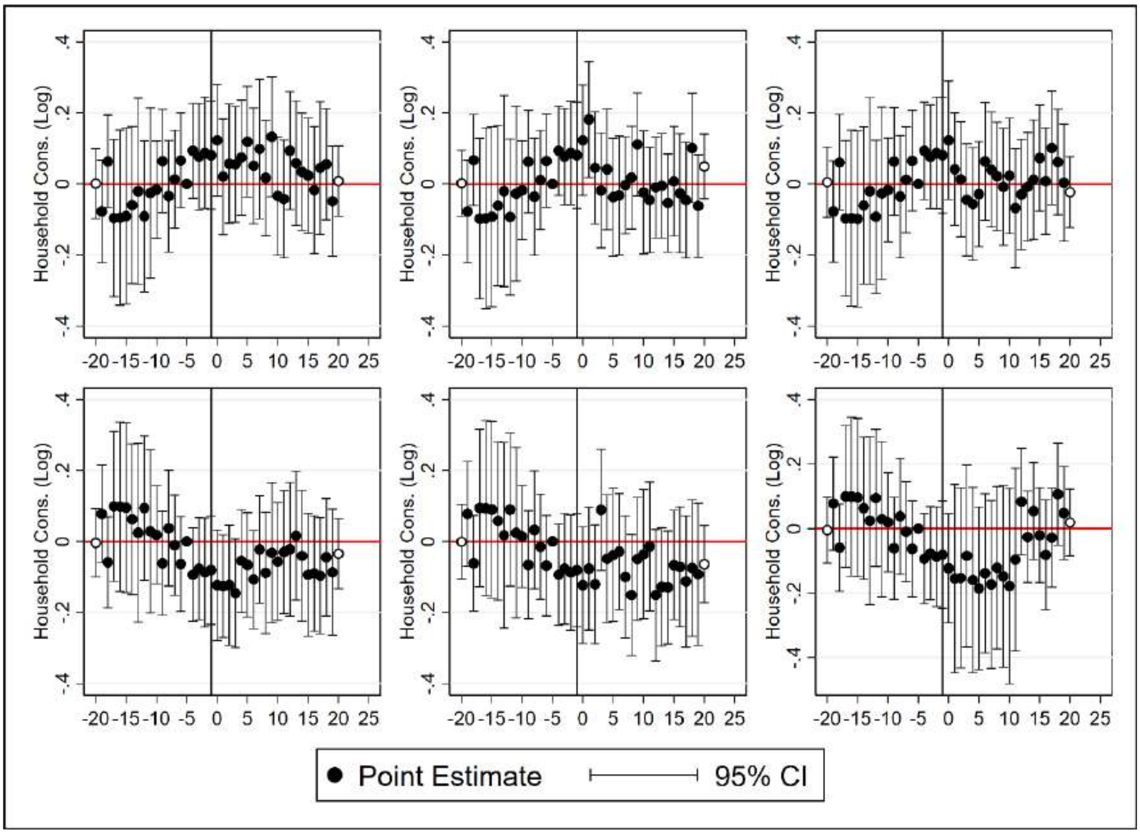


Figure 3-8: Event study results for household consumption in for group 1 in upper panel and group 2 in lower panel

3.2.5 Group-dependent effects: Participants with and without controllable EVs

Finally, we also investigate the differences between the EV and the PV groups that are unique to the German pilot. This is done by repeating the main regression model (cf. Table 3-3), but with a distinction between the two groups. For ease of exposition, the coefficients and confidence intervals are again presented graphically. Fitting to the previous results, there is a common pattern for nudges 1 and 2, but a different result for nudge 3. For the first two nudges, the EV group responds much more strongly than the PV group. Prosumers with a controllable EV get exposure to the nudges through a mobile application in addition to the common web portal. The results, therefore, indicate that the mode of nudging matters for the effectiveness of the intervention. For autarky, the effects are significantly different from zero for both groups, but only marginally different across groups. For self-consumption, however, the gap between the groups is much bigger and we have significance only for the EV group. The EV group increases self-consumption by 112%. Compared to the average effect of 2-3% over both groups, this is a sizable difference. The results thus suggest that the small average effects mix the null result for the PV group with sizable treatment effects for the EV group. For nudge 3 (opt-in), we would expect a similar pattern, given that the smart charging feature specifically targets the EV group. However, this is not the case. There are no significant differences between the PV and the EV group, as evidenced by the largely overlapping confidence intervals. The effects are negative for autarky and positive for self-consumption, so the direction of the effect matches the main result. The low activation rate of the opt-in is a likely explanation for this result. When only few households actively use the app, this nudge effect is not strong enough to drive the coefficient for the EV group as a whole. Heterogeneity in adoption probability and adoption timing is too big an obstacle for a clear identification in this context.

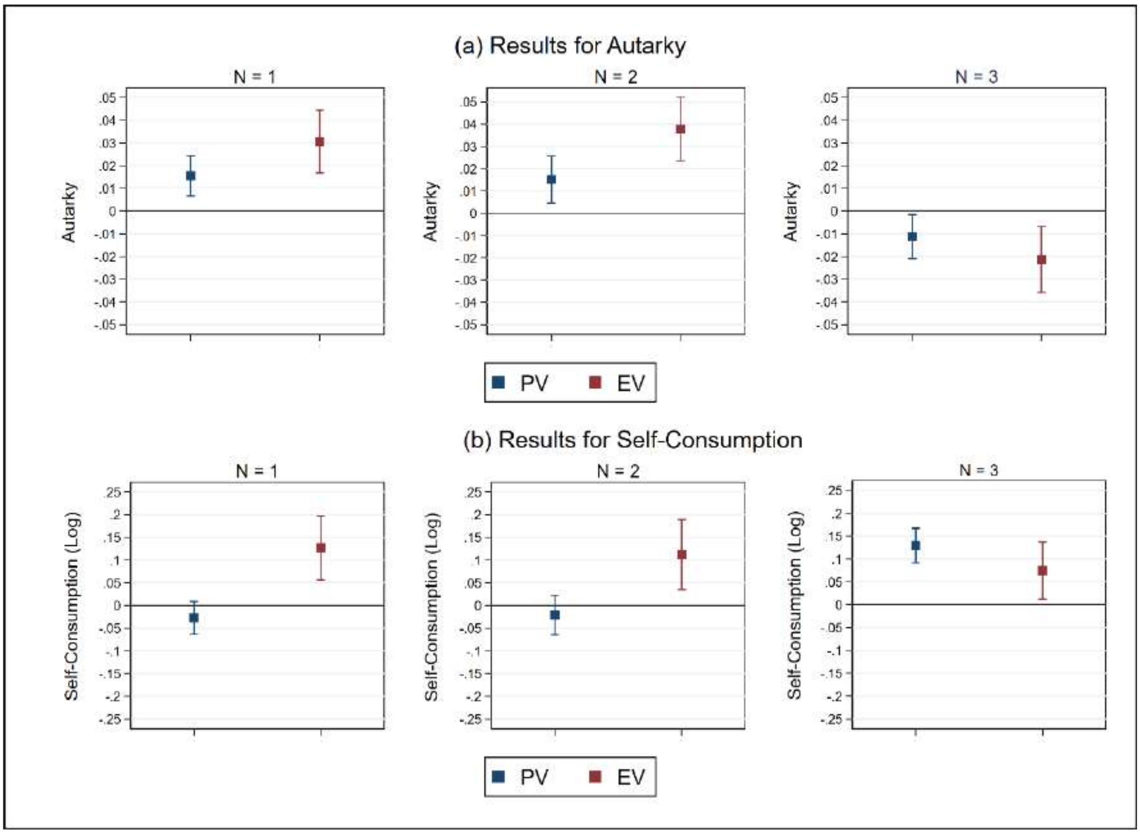


Figure 3-9: Sub-group analysis for self-consumption of EV and PV participants in group 1

For group 2, the heterogeneity analysis produces mixed results. The results for nudge 1 are the reverse of group 1. Again, the EV group is more responsive than the PV group, but here the EV group has a stronger *negative* coefficient. For nudge 2, there is no clear pattern and the confidence intervals for the EV group are too wide to allow for any definite conclusions. For nudge 3, the EV group responds more in autarky, but less in self-consumption. However, this last weak effect for the EV group is particularly interesting because it indicates that the negative average effect here is driven by the PV group, which did not even have access to the smart charging function. Hence, the surprising negative effect for the opt-in nudge is unrelated to the smart charging feature. We are unable to identify individual households that account for the negative effect in the PV group, but a one-by-one search of the descriptive patterns suggests larger heterogeneity in this latter period. There are at least three potential explanations. First, the accompanying survey data (see next section) indicates that there were several changes to household composition and equipment throughout nudge 3. Second, this is the only nudge that happened during heating season. We ensured that heat pump ownership is balanced across the sample, but it is not possible to fully rule out any influence of different heating technologies across households that were potentially underreported in the survey on which we rely for robustness checks. Third, nudge 3 was implemented in early 2023, which in Germany coincided with the launch of several policy packages to address the energy crisis, which may have spurred a larger cycle of behavioral change that overlays the relatively subtle nudging treatment with the energy report (the PV group only got this report, not the smart charging feature). Looking closely at the development of the raw data, there is indeed some evidence of a divergence between groups towards the end of nudge 3, when the second group is still treated. With these caveats in mind, the results for nudge 3 in group 2 are considered inconclusive.

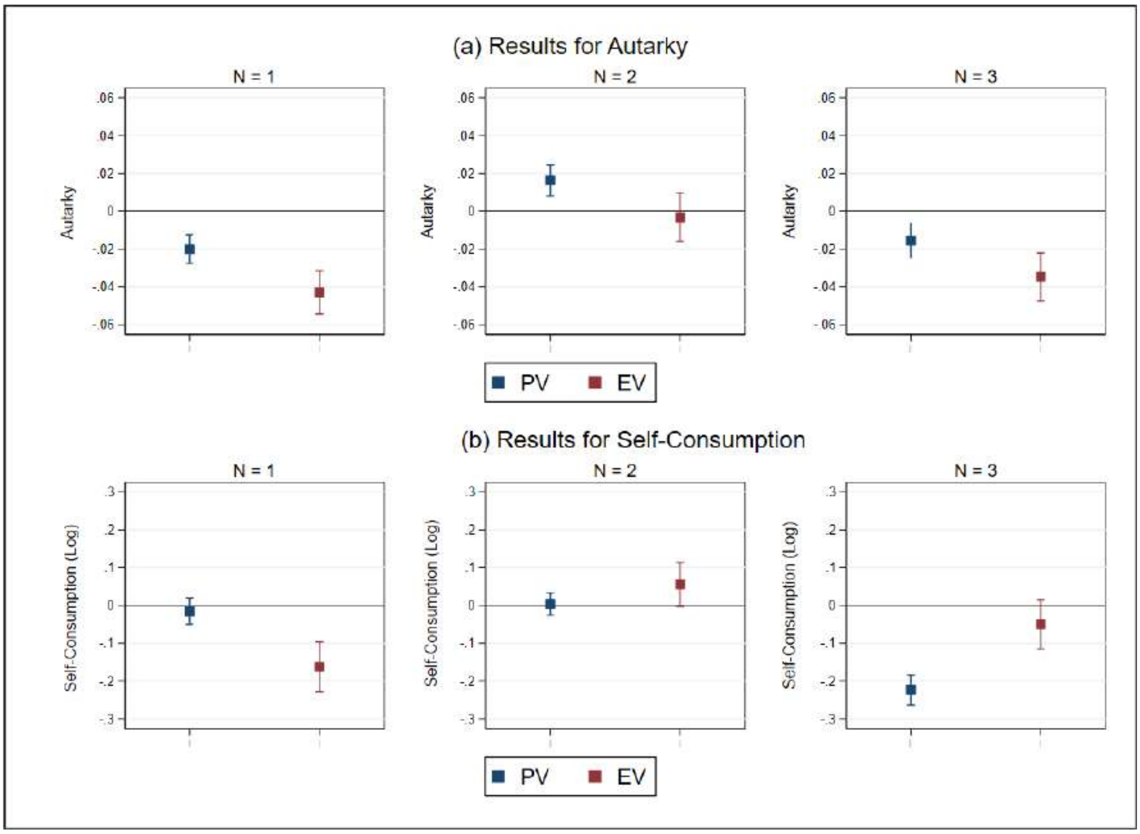


Figure 3-10: Sub-group analysis for self-consumption of EV and PV participants in group 2

For household consumption, the hypothesis for heterogeneity is not clear ex-ante. The smart charging feature should be unrelated to household consumption, unless there is an unobservable correlation between charging behavior and household consumption. The energy report was made available to both groups. Nevertheless, Figure 3-11 plots the heterogeneity to complete the analysis. In group 1 (upper row), there are no sub-group differences during the first two nudges – the confidence intervals overlap, and the coefficients have the same negative direction. In nudge 3, there is a positive effect for the PV group, but a negative effect for the EV group. This accounts for the average effect being close to zero in the main result (Table 3-3). It appears that the PV group slightly increased their household consumption – which is not in line with expectations and may reflect rebound effects. The EV group, however, shows a significant decrease in household consumption.

For group 2, there is again no evidence for sub-group differences during nudge 1. For nudge 2, it appears that the PV group does not respond, while the EV group increases household-consumption. This reveals that the positive effect in the main result is driven by the EV sub-group. For nudge 3, the results are the exact opposite to group 1. This raises concerns that there may be a general divergence between the groups after the start of 2023, which may bias the results. If household consumption changes in response to policy framework updates and electricity price changes after the turn of the year, the control group and the treatment group may adopt different trajectories. For example, some households invest in new energy-efficient technologies, others change their daily routines, etc. If there is a common pattern of behavior, the time-fixed effects can absorb this. However, if the responses are not evenly distributed across the two groups, this can lead to bias in the results.[39] With the small samples in the sub-groups, the results are generally sensitive to such effects. Overall, the results do not show a clear pattern across nudges and groups. It appears that household consumption is relatively more idiosyncratic across time than the two self-consumption indicators.

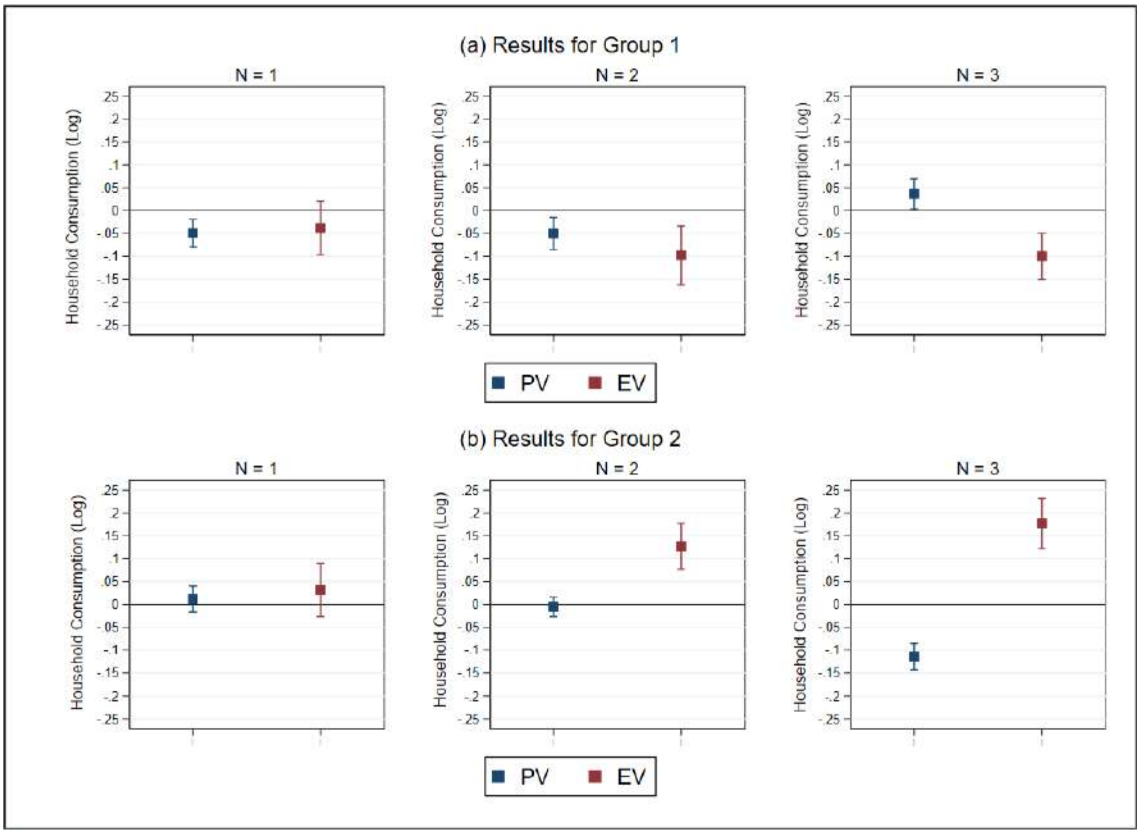


Figure 3-11: Sub-group analysis for household consumption in both groups

[39] In econometric terms, this constitutes a violation of the stable unit treatment value assumption.

3.3 Testing hypotheses on intention and motivation & further analyzes

3.3.1 Intention and motivation

The Figure 3-12 shows the development of intention and motivation in a bar chart. All mean values are greater than 3 and thus lie above the middle of the scale indicating a generally high motivation and intention to save energy and use more of the own PV electricity.

The intention to save energy increased significantly after Nudge 1 and 2. Interestingly, the intention to save energy significantly decreased after Nudge 3. This may be caused by the type of the nudging (default setting), which required no behavioral change from the households. Overall, the intention to save energy increased significantly from Nudge 1 to Nudge 3.

Intention to use more of one's own PV electricity increased significantly after Nudge 1 and remained approximately at this level after Nudge 2. The intention to use more of one's own PV electricity decreased slightly after Nudge 3 (for explanation see above). However, there was a significant increase in the wave 1-4 comparison.

Regarding the motivation to save energy, there was a significant decrease in motivation to save energy after Nudge 1. The motivation remained stable between Nudge 1 and Nudge 2 and slightly decreased after Nudge 3. Overall, there was a significant difference in the wave 1-4 comparison indicating that the motivation to save energy decreased from the baseline (before nudge 1) until the end of the intervention phase (after Nudge 3).

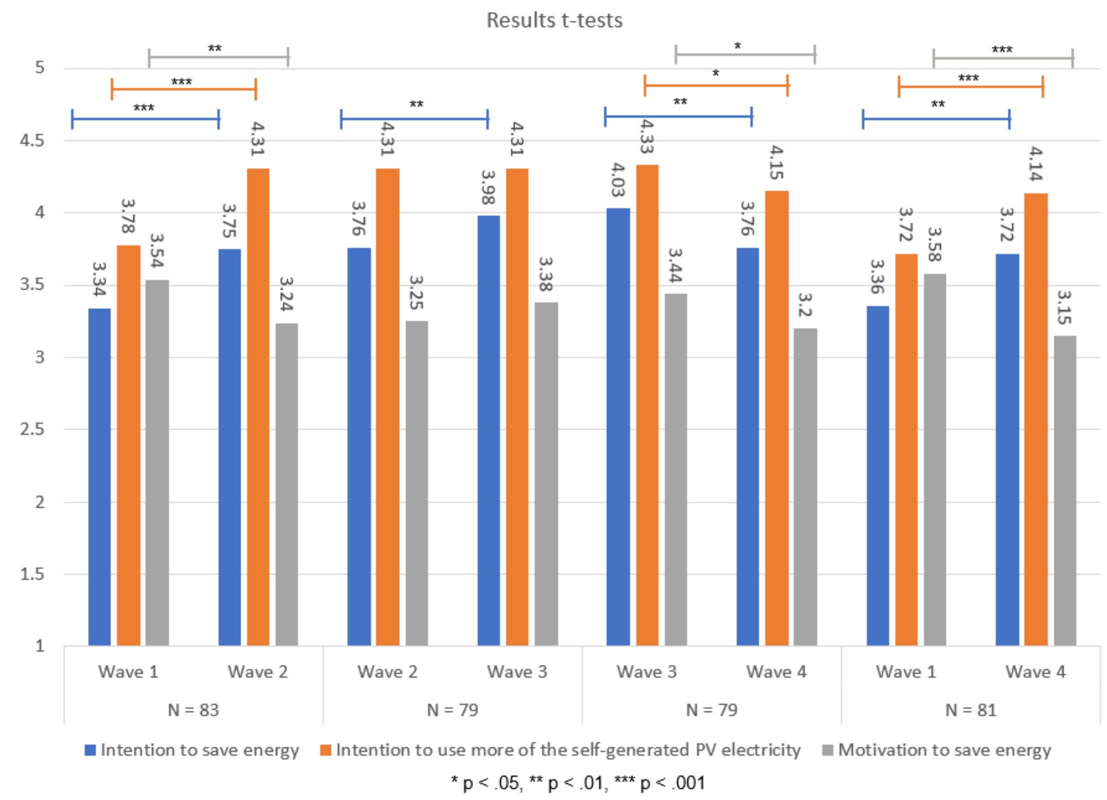


Figure 3-12: Intent to save energy across waves in the German pilot

3.3.2 Household changes over the nudging periods

Minor differences between both groups exist concerning their technical equipment and routine during the nudging periods. Four more participants own a stationary battery in group 2 than in group 1. Apart from that, both groups are similarly equipped concerning the installed PV capacity, the availability of EVs, wall boxes, heat pumps, and their technical dimensions (see

Table 3-6). During the nudging periods, the technical equipment changed for some participants. In particular, five participants became owners of an EV, and one participant purchased a heat pump. Four out of the six new ownerships concerned group 1 participants and happened during the nudging period 2. Four participants upgraded their PV capacity and one her battery storage (see

Table 3-6).

The routine of the participants changed during the nudging periods and between the groups (see Table 3-7). Group 1 reported being mainly at home for more days than group 2. Both groups spent most of their time at home during nudging period 1 (April till mid-July' 22), which can be associated with the Corona pandemic. Also, both groups reported increases in the number of people living in their house (i.e., mainly group 1 during Nudge 1, group 2 during Nudge 2).

Table 3-6: Technical equipment of both groups in the German pilot

	Mean or #	SD	Min	Max
Group 1 (n=54)				
Installed PV capacity [kWp]	7.94	3.18	3.64	19.2
Number of EV owners (controllable and non- controllable)	32	-	-	-
Number of heat pump owners	14	-	-	-
Number of battery owners	50	-	-	-
Battery volume [kWh]	6.73	1.65	4.59	10.37
Number of wallboxes	30	-	-	-
Installed wallbox capacity [kW]	15.08	5.44	11	22
Group 2 (n = 52)				
Installed PV capacity [kWp]	8.46	3.00	3.63	19.47
Number of EV owners (controllable and non- controllable)	32	-	-	-
Number of heat pump owners	15	-	-	-
Number of battery owners	54	-	-	-
Battery volume [kWh]	6.87	1.85	4.59	12.8

Number of wallboxes	30	-	-	-
Installed wallbox capacity [kW]	15.13	5.48	10	22

Table 3-7: Changes during the experiment in the German pilot based on the self-reports in survey data

	Baseline (n = 52)	**Nudge 1 (n = 45)**	**Nudge 2 (n = 42)**	**Nudge 3 (n = 41)**
Group 1				
Average of days being mainly at home per week	5.87	5.89	5.60	5.73
Share of participants using the web portal more than weekly	56%	76%	76%	83%
Share of participants installing and using the smart charging app more than weekly	2%	22%	18%	13%
Share of participants with an increase in electricity prices during the nudging period	-	20%	33%	41%
Other changes reported during the nudging period (Number of HH)	-	More people living in HH (3), new EV (2), new air conditioner (1)	More people living in HH (2), New EVs (3), new heat pump (1), new electric devices (4)	New EV (1), new PV (1)
Group 2	**Baseline (n = 52)**	**Nudge 1 (n = 45)**	**Nudge 2 (n = 51)**	**Nudge 3 (n = 46)**
Average of days being mainly at home per week	5.42	5.47	5.29	5.43
Share of participants using the web portal more than weekly	50%	67%	59%	67%
Share of participants installing and using the smart charging app more than weekly	9%	40%	35%	11%
Share of participants with an increase in electricity prices during the nudging period	-	20%	33%	52%
Other changes reported during the nudging period (Number of HH)	-	More people living in HH (3), new EV (1), new electric devices (2),	More people living in HH (3), new electric device (1),	More people living in HH (1), new EV (1), new

	disconnecting single rooms from electricity (1)	building renovation (1)	electric device (1), new PV (1)

3.4 Summary

Feedback and comparison nudges result in modest self-consumption increases, typically in the range of 3-4 percent. However, the opt-in setting of the smart charging app leads to a substantial 16 percent increase in self-consumption among active participants. It's worth noting that households with controllable electric vehicles tend to exhibit more pronounced effects compared to those without such vehicles.

Our findings strongly suggest the effectiveness of nudges in establishing new routines, especially when dealing with high-consumption technologies like EVs. To scale up nudges for other prosumers, we recommend implementing nudges that require minimal user interaction and energy literacy. Even among our self-selected and motivated participant group, interactions with the nudging tools were infrequent, and motivation showed a decreasing tendency.

4 Croatian pilot: Increasing self-consumption

In the following section about the Croatian pilot, one pilot-specific hypothesis is tested based on sensor data. The two general hypotheses on intention and motivation are tested based on survey data. As outlined in Table 4-1, we are able to confirm the pilot-specific hypothesis partially and the two general hypotheses. The section references in the table guide the reader to the analysis based on which the hypothesis is tested. Additionally, we provide an analysis of the impact of the Croatian regulatory framework on self-consumption.

Table 4-1: Tabular summary for the Croatian pilot

Hypotheses		Based on	Outcome	Section reference
HR1	Nudges are effective in increasing the self-consumption of participants.	Within-subject with sensor data (1 or 2 week(s) before & during nudge)	✓	4.2.2
		DiD with sensor data (daily aggregation, wash-out vs. nudge)	–	
All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	✓	4.3.1
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	✓	4.3.1

4.1 Pilot-specific research design

4.1.1 Hypotheses

Similar to the German pilot, the participants in the Croatian pilot are prosumers with rooftop-photovoltaic who aim to increase their self-consumption based on nudges. Due to challenges in the acquisition of participants and bottlenecks in the supply of sensors and technicians for their installation during COVID-19, the sample increased during the three nudging periods. During the third nudge, we reached a sample of 82 participants. We test the following hypothesis for these participants:

HR1: The nudges are effective in increasing the self-consumption of participants.

A particularity in the Croatian case is the regulatory framework that draws a distinction between "self-consumption" and the "final customer with own production" model, depending on whether the amount of produced electricity exceeds the amount of consumed electricity. The system and its potential implications for household behavior are detailed below. While this was not the focus of the experiment, the regulation has emerged as a major driver of behavior that may override the nudge effects or yet increase the incentives. Nudge 2, in particular, provides participants with information regarding their standing regarding the policy.

The potential effects of this regulatory framework are summarized by the following two hypotheses:

HR2: Participants (who run the risk of producing more electricity than consumed on an annual basis) increase their consumption.

HR3: Participants (who run the risk of producing more electricity than consumed on an annual basis) curtail their production.

We address the effects here briefly where they concern the nudging interventions, a more detailed exposition is placed in Deliverable 1.3.

4.1.2 Analytical strategy

Given constraints in the sample size due to a delayed roll-out, the difference-in-differences strategy applied to the German pilot was deemed inappropriate for the Croatian sensor data. After analyzing the available data descriptively, we instead chose a regression discontinuity design (RDD) focusing on the days around the nudge introduction. This delivers the short-term effect of the nudge from a simple before-after comparison for the treated group. This circumvents problems with the baseline and the low number of households in group 2 initially. The assumption is that more profound differences across groups that build over time and the changing composition of the treatment and control group are not relevant in the short run. To increase power to a sufficient level, we stack the two groups and evaluate a single effect across both cohorts for each nudge. We also check the individual groups, the balancing of co-variates, and alternative functional forms and time frames with this method. The disadvantage of the method, however, is that it does not employ a control group because it relies on the assumption that the introduction of the nudge is the only thing that changed in the long run. Sudden weather changes would, therefore, be a problem if this happened exactly around the day of implementation. This seems like an acceptable trade-off against the gain in sample size, but we nevertheless explore whether this assumption is reasonable by cross-checking weather data and by estimating a short-run difference-in-difference method that suffers from power problems but can at least validate whether the RDD strategy is internally valid.

For the survey data, we follow the same strategy as the other pilots by focusing on intention and motivation to save. We additionally evaluate the questions pertaining to the above policy framework in more detail as a country-specific factor.

For the app data, we will check the adoption of the app and the activity levels over time. This is planned as a descriptive analysis as for the other pilots, with a possible extension to a regression analysis on the effects for active participants as in the German pilot.

4.1.3 KPI & data

The data structure and KPIs are the same as in the German pilot. We run regressions to evaluate the effects on self-consumption and autarky. Because the Croatian regulations provide additional incentives to change total consumption, we add this as a third outcome variable and report the results when there are additional insights from this added analysis.

We investigate individual households' behavior more deeply by combining the survey and the sensor data. Considering the power constraints from the small sample size, this additional analysis is planned as a descriptive evaluation. This also includes a descriptive evaluation of the differences between the three main locations of the participants, namely Zagreb, Osijek, and Varaždin.

4.1.4 Nudging interventions

We introduced a new app that presents information on the household's consumption and production to convey the nudges. These impulses aim to motivate households to shift their consumption to hours of self-generation or increase it during hours of self-generation. While the first nudge approaches the participants with empathy raising messages on energy poverty and pollution, the second nudge visualizes the consumption and production in a simple, tangible manner, and the third nudge invites participants to set and enforce targets about their self-consumption (see Figure 4-2).

		Nudge 1: Instigating empathy		Nudge 2: Feedback		Nudge 3: Target setting	
Group 1	Baseline	Treated	Control	Treated	Control	Treated	Control
Group 2		Control	Treated	Control	Treated	Control	Treated
Timeline	Jan - Jul `22	Aug - Sep `22	Sep - Oct `22	Nov - Dec `22	Dec`22 - Jan `23	Feb - Mar `23	Mar - Apr `23

Figure 4-1: Experiment outline for the Croatian pilot

In particular, the first nudge sends a message from the perspective of the ones who suffer from the consequences of not performing in energy conservation - we would like to include a perspective from people living in energy poverty. These are messages such as: "You used X% more energy than you produced yesterday. This amount could help heat the home of an energy-poor family in the winter." These messages are reinforced by an illustration of the Earth with different levels of pollution depending on the household's CO_2 emissions.

The second nudge provides feedback on the self-consumption level. It also offers an overview of the aggregated consumption and production amount that is relevant for the regulatory assessment that determines the requirements for feeding self-generated electricity into the grid. To avoid the additional requirements when the amount of production is higher than the amount of consumption, households can start to consume more self-generated electricity or curtail their photovoltaic plants.

The third nudge on target setting asks the participants to set a target for their level of self-consumption. It also visualizes the delta for reaching the target.

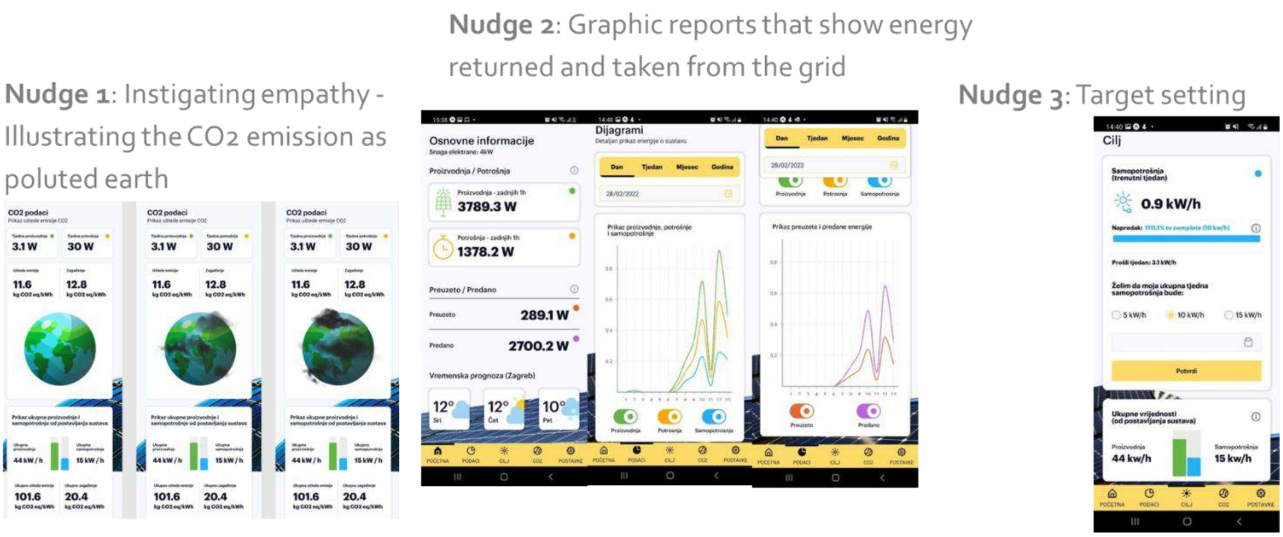


Figure 4-2: Nudging interventions of the Croatian pilot

4.1.5 Croatian policy framework for prosumers

From 2021 until the end of 2023, the Croatian legal framework consisted of two distinct regulatory models: the "self-consumption" model, which applied to households and public institutions, and the "final customer with own production" model, which encompassed all other customer categories but can also include households. These models were established by the Law on Renewable Energy Sources and High-Efficiency Cogeneration (Article 51). For household PV systems, a household transitioned to the "final customer with own production" model if they exported more energy to the grid than they imported in a given year. Under this billing approach, surplus energy that is not self-consumed on-site is bought by suppliers at a minimum of 90% of the user's average electricity price. Unlike the "self-consumption" model, which allows netting within a month, this model does not offer any netting. This significantly impacts investment returns, typically resulting in a 30% decrease. Additionally, the status switch comes with an increased compliance burden that also makes it undesirable from a non-monetary perspective. The surplus in the regulation is defined as grid-in minus grid-out. In practice, the policy leaves prosumers two margins of adjustment: increase energy consumption (i.e., increase grid-in) or reduce the PV plants' production (i.e., grid-out). We analyzed the response to these regulatory models in Kesselring et al. (2023)[40].

4.2 Testing pilot-specific hypotheses

4.2.1 Descriptive statistics

The Croatian pilot started later than initially planned due to supply shortages that caused delays in the installation of the required hardware. Figure 4-3 plots the number of households from which data are available over time. The solid black lines denote the start of the 3 nudges, the dashed lines the switch of the treatment from group 1 to group 2. The red dashed line is the first time the sample size exceeds 30, which was deemed the minimum number required for regression analysis. The graph shows why a full difference-in-differences strategy is not appropriate for Croatia – the sample is very small, and especially group 2 only starts growing in the summer. Adding in the volatility during vacation time in August (also in the German pilot), the evaluation of nudge 1 is hardly feasible. Two more takeaways from the plot emerge. First, the fluctuation of the sample size at daily aggregation is still apparent, which may come from data gaps due to technical reasons or deliberate behavior by households. This is not a threat to the analytical strategy per se, but compounds concern about the small sample size. Second, a comparison across the different interventions is not directly possible given the ongoing changes in the composition of the sample, i.e. new households keep being added and some households do not send data every day. If the households added later differ systematically from the early adopters, then the comparison draws on two different subsets. A direct comparison requires the assumption that each added household is representative of the overall sample. We, therefore, choose to evaluate each nudge separately before drawing any conclusions about their relative effectiveness.

[40] Kesselring, Anne; Pelka, Sabine; Svetec, Erica; Nad, Lucija; Seebauer, Sebastian; Skardelly, Sara; Preuß, Sabine: Slashing the surplus – how prosumers with smart metering respond to regulatory restrictions on self-consumption in Croatia, Conference Proceedings, BEHAVE 2023

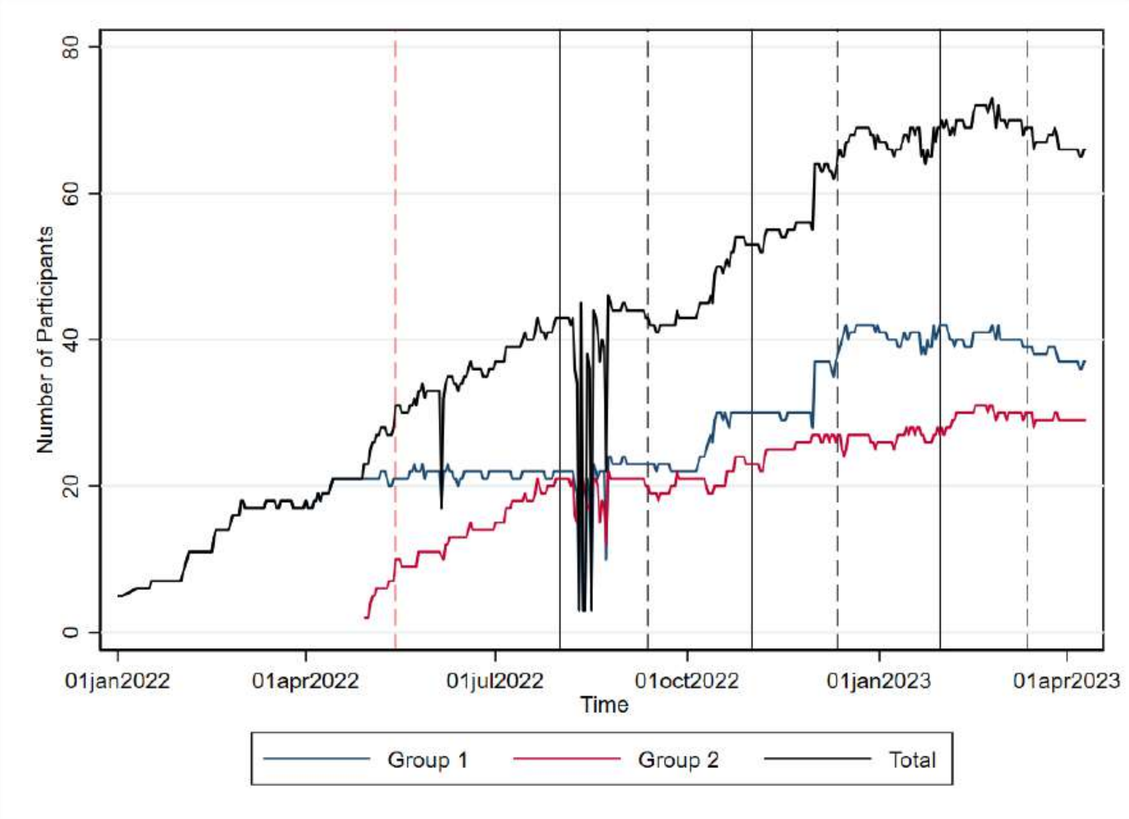


Figure 4-3: Number of participants based on transmitted data by group over time in HR pilot's sensor data

Figure 4-4 plots the two main KPIs autarky rate (0 to 1) and self-consumption (mean hourly value during 24-hour period), comparing the two groups. During the baseline (N = 0), the two groups start to converge during nudge 1 (N = 1), following a common trend that fluctuates with the weather in the short-run and also has a seasonal component. Relative to the German pilot, the drop in the winter months is shorter and less pronounced, as expected due to the different climate. These first results confirm the impression from the previous graph: the baseline and (to a lesser extent) the first nudge are not suitable for a difference-in-difference analysis, but the series does stabilize sufficiently to allow the use of estimation strategies that focus on the short-run fluctuation. For data preparation, we also looked at the distribution of values in the sample beyond the mean. Here, it becomes apparent that there is a wide spread of values in the Croatian sample. While some households are very close to complete autarky from April to October, there is also a substantial number of observations with very low or zero autarky. The high heterogeneity in the sample suggests that households in the Croatian pilot pursue very different strategies of operating their PV plants throughout the year. This observation may be linked to the unique policy setting in Croatia (see previous section), but it may weaken the potential for effective nudging. Households that are already close to autarky may have little room to optimize self-consumption further. Households that do not make use of their PV plant continuously may not be reached by the nudge at a suitable time and therefore also exhibit a weaker behavioral response.

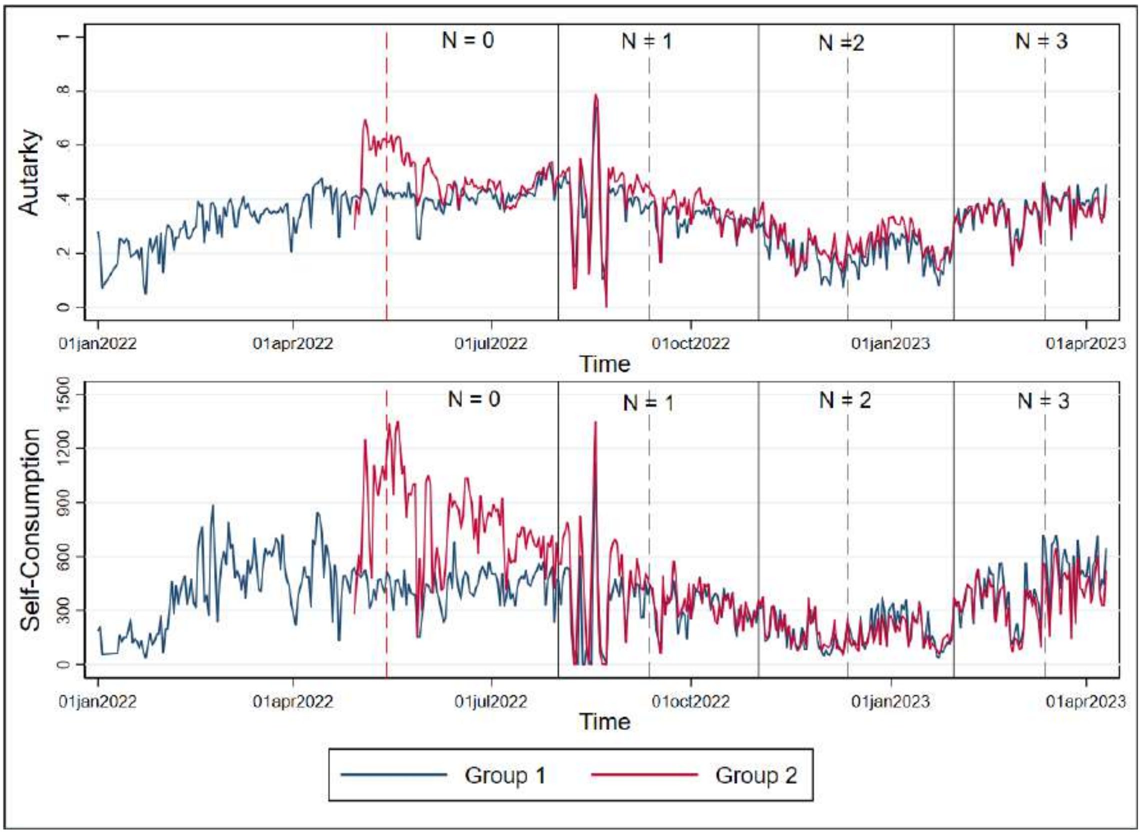


Figure 4-4: Indicators of self-consumption by group over time

Table 4-2 provides descriptive statistics for the two groups over the sample period. The lower number of observations (N in last column) is a consequence of the late start for group 2. Otherwise, the groups are shown to be comparable, which indicates that a strategy involving a group comparison is feasible for the later nudges, with the stipulation that the baseline (N=0) is not used as the reference.

Table 4-2: Summary statistics by group

Group 1	**Mean**	**SD**	**Min**	**Max**	**N**
Consumption [Wh]	1004.91	1192.72	0.04	18218.16	242229
Self-Consumption [Wh]	331.7	789.37	0.00	9955.7	242229
Autarky (0 to 1)	.3	.38	0.00	1	242229
Group 2	**Mean**	**SD**	**Min**	**Max**	**N**
Consumption [Wh]	969.42	1123.36	0.00	13971.55	150196
Self-Consumption [Wh]	335.68	796.64	0.00	9804.41	150196

Autarky (0 to 1)	.3	.38	0.00	1	150196

Notes: *Descriptive statistics for estimation sample from January 2022 to June 2023 at daily aggregation. Self-consumption is the difference between total consumption and output to grid. Autarky rate is the ratio of self-consumption to total consumption.*

4.2.2 Nudge effects

Figure 4-5 shows the results of the RDD in graphical form. The day on which the nudge is activated is marked as zero, and the timeframe is restricted to a period of 7 days before and after. The grey dots display the mean outcome on that day, and the black lines are fitted regression lines from the data analysis. The treatment effect can be seen from the discontinuity between those fitted lines where they meet at the red cut-off line. The regression lines allow for differences in slope (i.e., the short-run trend leading up to the nudge does not need to be the exact same as the short-run trend during the nudge). Small, continuous changes in weather would be an example for a slope difference. However, if there are large discrepancies between the slopes on both sides, this is a warning sign that the two sides may not be comparable. Note that self-consumption in the lower panel is reported on logarithmic scale.

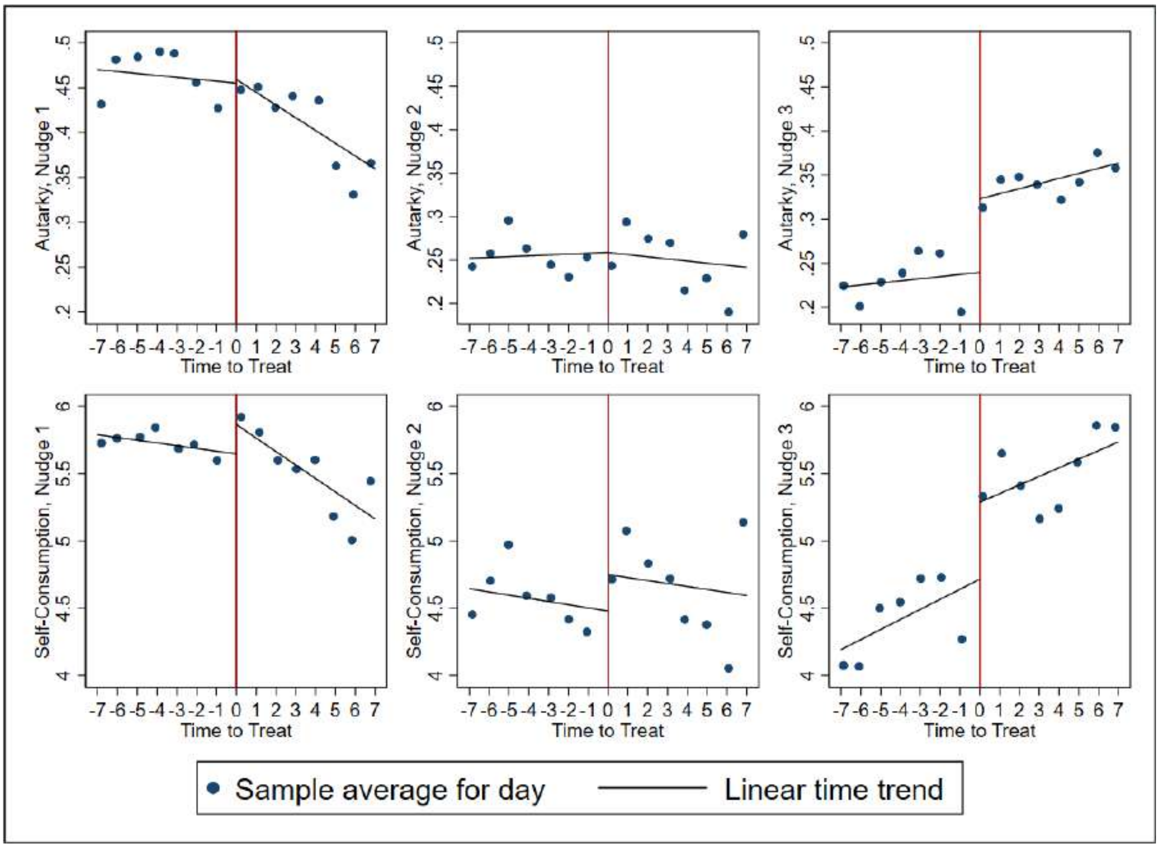


Figure 4-5: RDD for time window of 7 days

For nudge 1, there is a small, upward level break at the cut-off. This would point to a positive treatment effect, but the slope break is so severe that the estimate is hardly credible. The result provides further confirmation that given the very small sample, an evaluation of nudge 1 is not feasible with appropriate statistical methods. For nudge 2, the results stabilize and indicate that there is no effect autarky, but a small positive treatment effect for self-consumption. The slopes match much closer than for nudge 2, which would be expected given the larger sample size (626 for nudge 1 versus 844 for nudge 2). For nudge 3, there is a clear discontinuity at the cut-off. This is interpreted as a positive treatment effect: both autarky and self-consumption increased during the nudging intervention (sample size: 1051).

The visually represented effects can also be quantified as a coefficient with a standard error and significance level. This is reported in Panel A of Table 4-3. The results are reported for the KPIs of autarky in column 1 and self-consumption in column 2, for further insights, total energy consumption is added in column 3. The effects for nudges 1 and 2 are insignificant across all specifications. For nudge 3, the estimate on autarky indicates an improvement of 8.3 percentage points. For self-consumption, the coefficient corresponds to 224 Wh when evaluated at the sample mean of 391 Wh during this period (point estimate of 0.573 is in logarithms), which is a very strong effect. Interestingly, total consumption increases at the same time. The coefficient of 0.253 indicates a smaller percentage change, but in fact corresponds very closely when converted from log-points to Wh at the respective sample mean: 249 Wh. The estimates therefore indicate that nudge three had a small, positive effect on the autarky rate that is driven by a simultaneous increase in both self-consumption and total energy consumption.

The results are robust to the use of a different bandwidth (time window). Figure 4-6 shows the results for a period of 14 days before and after the nudge. The specification is otherwise the same as above. For nudge 1, the time series is volatile especially in the second week of the intervention, but given the very small sample, we do not attach weight to this observation. For nudges 2 and 3, the results prove to be robust to the choice of the different time window. This also holds for windows of 10 or 21 days (not shown here).

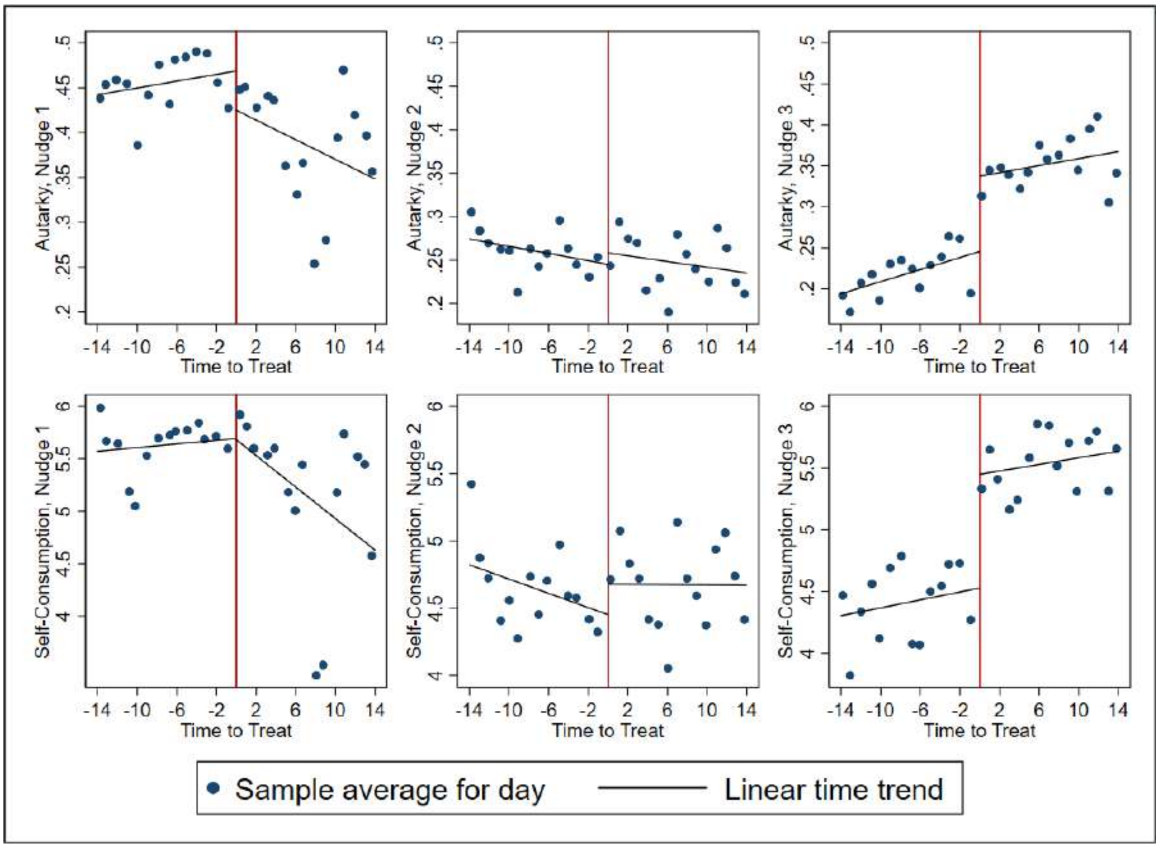


Figure 4-6: RDD for time window of 14 days

The critical assumption behind the regression discontinuity design is that all other factors that determine self-consumption do not have any sharp breaks at the same time that the treatment in introduced (in technical terms: relevant confounders are balanced around the threshold). The main threats to the analytical strategy that were previously discussed are weather and the energy crisis. It is important to note that the short-term focus of the regression discontinuity design sidesteps the long-term effects of the energy crisis because it is highly unlikely that such broader shifts will take effect precisely within the one week before and after. For the case in question, the main concern is, therefore weather variation. The Croatian pilot is spread across three geographical areas, and since we do not have sufficiently precise weather data for all observations, we are unable to control this precisely with the current data. We checked with weather data from the Copernicus Climate Change Service (C3S), and there does seem to be a weather swing around the time of the implementation of nudge 3 for group 1. If this is influential, we would overestimate the importance of the nudge in the estimated treatment effect.

Figure 4-7 displays the weather development. The specification is the same RDD, but the outcome variable is radiation (in J/m², divided by 10^5 for clarity). The plots in the upper row refer to group 1, the plots in the lower row refer to group 2. Two issues stand out. First, the weather for group 2 is not available during nudges 2 and 3. For nudge 2, December 2022 is missing. For nudge 3, the data are not available beyond March 1, 2023 (last checked on October 1st, 2023). These series are highlighted in red. The variable is coded as radiation = 1 because missing values are not handled by the RDD statistical package. Second, there is a visible break in the series at the implementation time of nudge 3 for group 1 (top right of plot). This is a coincidence that could not be foreseen, but it is a potential confounder to our results. The missing weather data in the lower row prevents us from controlling for weather, which would otherwise be a simple correction. For group 1 alone, the sample size is too small for meaningful inference, which is why we had stacked the two groups in the RDD.

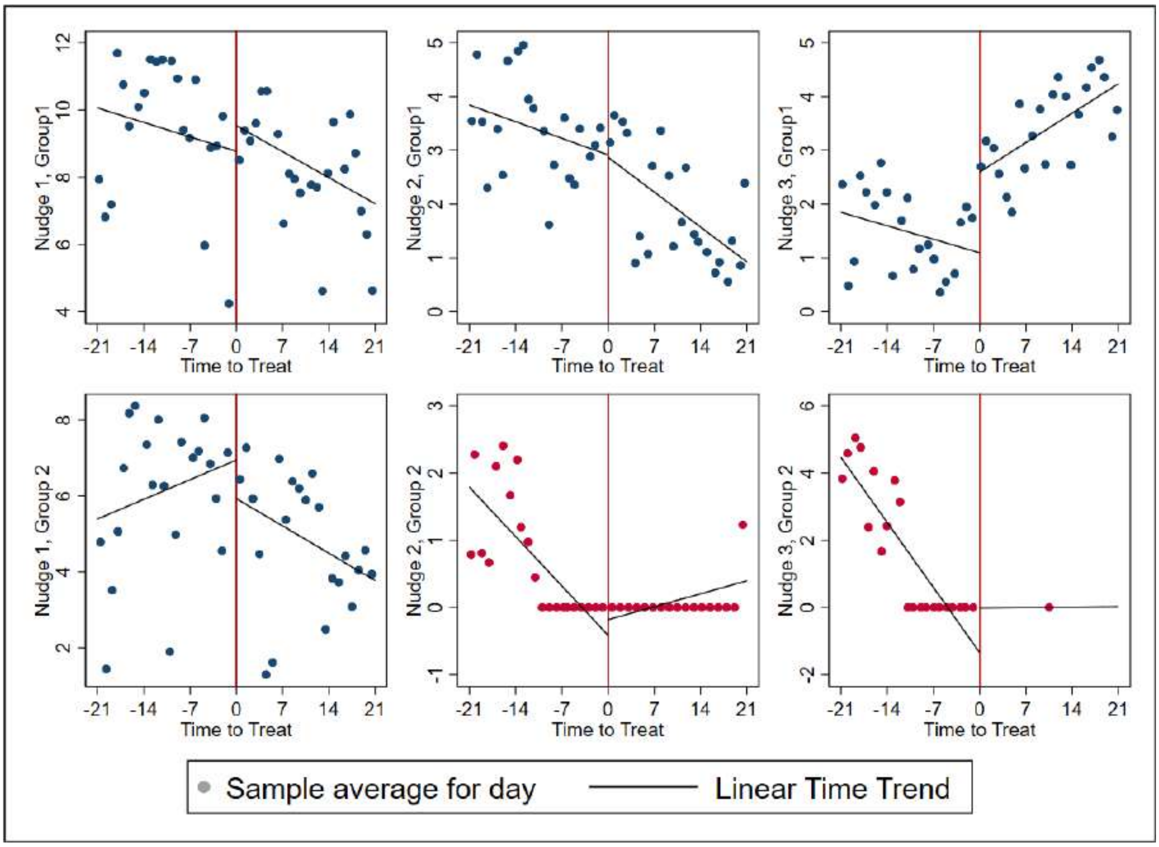


Figure 4-7: Covariate test for radiance in both groups

In a robustness check, we therefore also estimate a short-run DiD design that uses the wash-out period preceding each nudge as a baseline. This approach has two disadvantages relative to the regression discontinuity. First, it is strictly valid only for group 1, because there is only a washout at the conclusion of each nudge. Second, the need to focus on a short-run window severely limits the sample size. It will be difficult to ascertain whether any insignificant results are due to a lack of power or represent a true null effect (in technical terms: high risk of type II errors). With these limitations in mind, we consider the short-term DID mainly as an additional robustness check that complements the main results from the RDD.

Nudge 1 is included for completeness in Panel B of Table 4-3, but with less than 20 households in each group, the reported negative estimates are not credible and would have to be cross-checked with a larger sample first. For nudge 2, it emerges that the effect on the three outcomes differs strongly: there is a small decrease in autarky, a large decrease in self-consumption, and a strong increase in total consumption. At first glance, these results seem counterintuitive because the hypothesis was an increase in the KPIs. However, further context with the Croatian policy matches with what we find here. Nudge 2 occurs late in 2022, so households must balance their surplus by year-end to avoid the switch of treatment status. The nudging intervention also includes a screen dedicated to the policy, which gives the treatment group a clear overview of where they stand. There are two ways of reducing surplus: turn down the PV generation (reduction in self-consumption) or add additional devices/switch to heating with electricity (increase in total consumption). The survey indicates that households did indeed take action in both directions. With this background, the estimated coefficients can be rationalized. We argue that the average effects are dominated by those users that respond to the policy by either reducing self-consumption through intentional PV curtailment, or by increasing energy consumption when they employ additional sources of electricity.

The end results may not be in line with the initial objective of the nudging intervention, but they speak to a co-benefit in the program: transparency. From discussions with the implementation partner at ZEZ and the insights from the Croatian national event, it became clear that the participants are highly concerned about the status switch if they ran a surplus. The information provided to them in the app allowed for better tracking and adjustment to avoid a surplus. The particular focus of nudge 2 and its timing to the end of the year made the policy the primary concern and triggered individual behavioral responses that likely prevailed over the relatively small behavioral change from the intuitive nudge.

For nudge 3, the policy urgency disappears because it happens in the beginning of 2023. Here, the main objective is to cross-check the results from the RDD given concerns about the weather. We find a positive effect on autarky. The magnitude of 3.9 percentage points is substantially smaller than the RDD result in Panel A. Similarly, the coefficients for self-consumption and total consumption are also positive, but substantially smaller in size than in the RDD. Both of these outcomes are no longer statistically significant. These results indicate that the positive effect for nudge 3 generally holds, but that weather effects contribute. The direction of the nudge effect is confirmed, but the effect size was likely overstated with the RDD. Notably, the effect size with the short-term DiD is now comparable to German pilots' feedback and comparison nudge.

Panel C adds the results for group 2. For nudge 2, we estimate negative coefficients for all outcomes, but it is important to caution that the calculations use group 1 as a control group, and the monitoring of the surplus by group 1 is likely to continue until the end of the year after the initial realization. Hence, the interpretation would be that group 2 responded less than group 1, which is a weak result given that we expect heterogenous responses to the policy across households. The sample is too small to provide meaningful sub-group analysis beyond the two-group comparison to explore this further. We therefore consider the response of group 1 as the best available evidence for the policy effect. For nudge 3, there are no significant effects for any outcome in group 2. This again indicates that this group did not develop any differently than group 1 during the intervention, which could be indicative of learning effects by group 1 that persist over time, but this hypothesis cannot be conclusively evaluated from the available data.

Table 4-3: Full list of coefficients and p-values for the Croatian pilot

	(1) Autarky	(2) Self-Consumption	(3) Total Consumption
		Panel A: RDD Coefficients	
Nudge 1	0.00401	0.218	0.0646
	(0.10)	(1.33)	(0.54)
Nudge 2	-0.000130	0.271	0.124
	(-0.00)	(1.40)	(0.82)
Nudge 3	0.0833***	0.573***	0.253**
	(2.80)	(3.82)	(1.98)
		Panel B: DiD for Group 1	
Nudge 1	-0.0285**	-0.0122	-0.00634
	(-2.23)	(-0.20)	(-0.11)
Nudge 2	-0.0529***	-0.227**	0.151**
	(-3.03)	(-2.26)	(2.34)
Nudge 3	0.0378***	0.0809	0.0768
	(3.43)	(0.92)	(1.29)

Panel C: DiD for Group 2

Nudge 1	0.00987	-0.0320	0.0181
	(0.84)	(-0.30)	(0.29)
Nudge 2	-0.0332**	-0.200**	-0.129**
	(-1.99)	(-2.02)	(-2.01)
Nudge 3	-0.00482	0.0225	0.0119
	(-0.37)	(0.32)	(0.21)

Notes: *Results in Panel A are obtained from regression discontinuity in time using a window (bandwidth) of 7 days and a linear time trend on either side (parametric 1st degree polynomial). Dependent variables autarky (0 to 1 ratio), self-consumption, and total consumption (log-transformed). Panels B and C show results from short-term difference-in-differences design using the 7 days prior to intervention as the baseline (common for both groups), and the first 7 days of the intervention (group-dependent) Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels:* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.2.3 Group-dependent effects: Descriptive analysis for different locations of the pilot

With the small sample, further sub-group analysis with a regression-based approach is not suitable in light of the existing power constraints. However, the Croatian pilot has additional information in the sensor and survey data that can be analyzed descriptively. We explore the development over time in the energy-related variables for each of the three cities to assess whether there are important sources of heterogeneity within the country. This analysis is especially valuable for the Croatian pilot, because the participants are prospective members of local energy communities.

Figure 4-8 shows the mean self-consumption levels (left panel) and the mean total consumption (right panel) for the three cities over time. We focus on the mean because the sums by city change with the number of households. The data are aggregated monthly for better visual representation. As the sample grows, the indicators quickly converge and then follow a very similar time path. During the holiday period in August, total consumption drops in all three cities, while self-consumption stays relatively stable. This is explained by efficiency in production during sunshine hours that bring up self-consumption, while consumption drops to zero when households are gone on vacation.

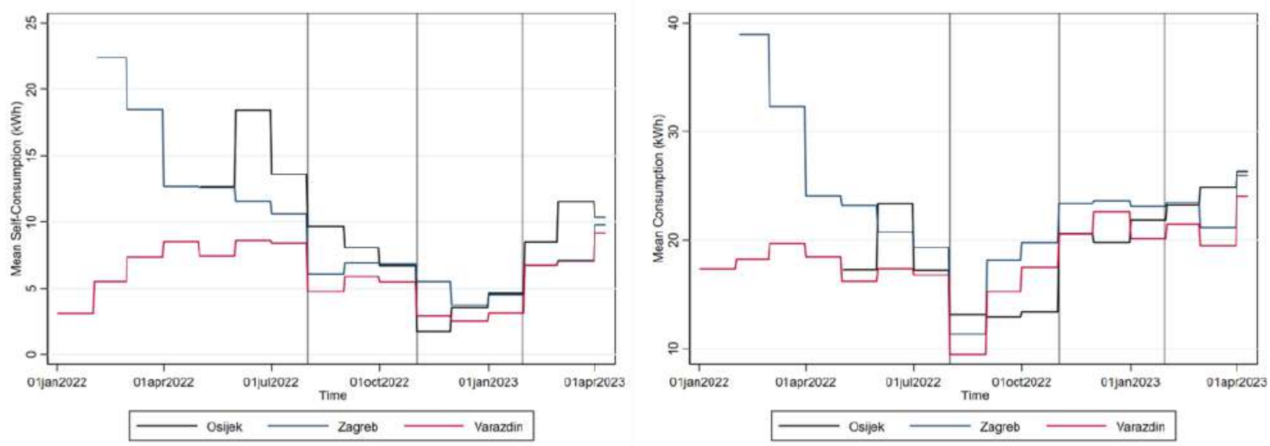


Figure 4-8: Self-consumption (left) and consumption (right) for three pilot locations in Croatia

Figure 4-9 complements the consumption patterns with statistics on grid-out (energy returned to the grid) and grid-in (energy taken from the grid). The trends are again similar, but there are some visible differences across cities. Osijek has a slightly higher grid-in during summer and a substantially lower grid-out during the fall. Zagreb households are on an increasing trend for grid-out after the turn of the year. Overall, the descriptive analysis supports the view that the cities develop in parallel. For policy interventions, it appears reasonable to extrapolate across locations when it comes to prosumer profiles.

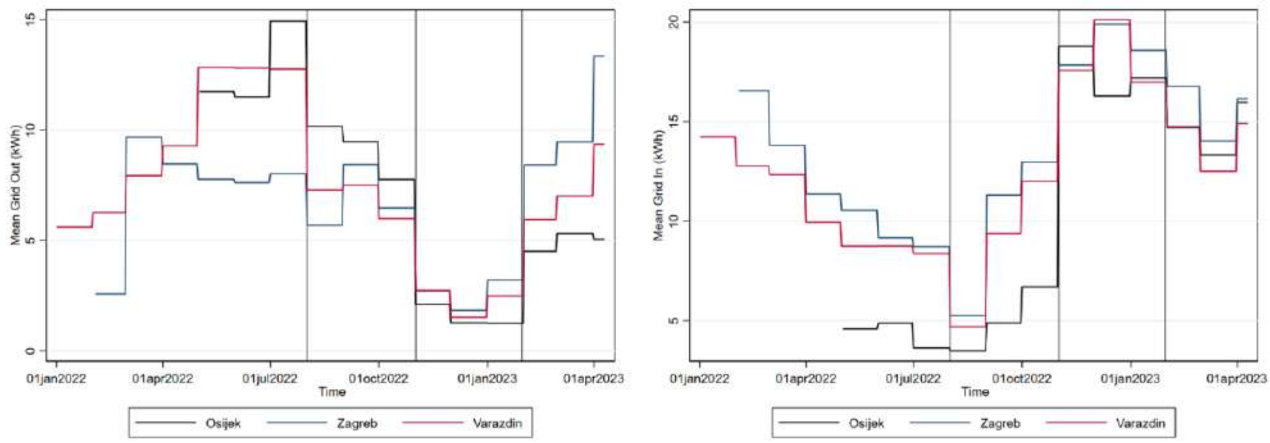


Figure 4-9: Grid-in (left) and grid-out (right) for three pilot locations in Croatia

4.3 Testing hypotheses on intention and motivation & further analyzes

4.3.1 Intention and motivation

Figure 4-10 shows the development of intention and motivation in the form of a bar chart. All mean values are greater than 3 and thus lie above the middle of the scale. While no significant changes in the motivation to save energy exist, the mean scores for intention to save energy to save energy moved slightly up and down over the course of the waves. Comparing wave 1 with wave 4, both values had increased. The intention to use more of one's own PV electricity decreased significantly in waves 2 and 3, then remained at this level in wave 4. Comparing the initial intention (wave 1) with the final one (wave 4), a significant decrease is shown. We associate this negative tendency with the Croatian policy framework for prosumers. As explained on page 63, if households produce more than they consume, they lose their self-consumption status and have additional requirements for feeding in their self-generated electricity. Nudge 2 increased the transparency on their ratio of consumption and production. This led to the decreased intention in wave 3 after nudge 2. We disentangle how participants responded to this policy framework in the following. You can also find this analysis in Kesselring et al. (2023)[41].

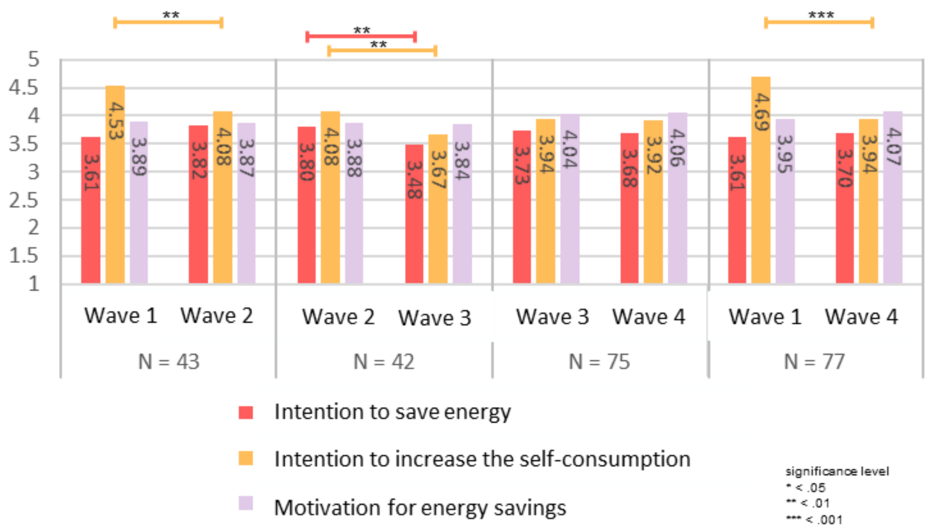


Figure 4-10: Intent to save energy across waves in the Croatian pilot

4.3.2 Response to policy framework for prosumers in Croatia

We asked participants about their electricity consumption behavior through the survey. Participants reported their energy consciousness (one item), electricity consumption, and PV self-consumption. Given the policy setting playing a big role at the turn of the year, we conduct additional analysis for survey waves in fall/winter 2022 (wave 2) and April 2023 (wave 4). Specifically, for the electricity consumption and their PV self-consumption, we implemented questions on the intention to save electricity (three items, Cronbach's alpha = .92 and .90, in wave 2 and 4 respectively) and the intention to use more PV energy (three items, Cronbach's alpha = .90 and .93) as well as their expected increase in electricity consumption (one item) and expected increase in self-consumption (one item). Thus, we collected data on five variables for electricity consumption behavior. The descriptive statistics of these variables are displayed in Table 4-4. We also examined the correlations between these variables and found two positive correlations in wave 1 and 2, namely between the intention to save electricity and the intention to use PV energy (wave 1: $r = .62$, $p < .001$ and wave 2: $r = .66$, $p < .001$) as well as between the expected increase of consumption and self-consumption (wave 2: $r = .84$, $p < .001$ and wave 4: $r = .71$, $p < .001$). In wave 4, we additionally found a small correlation between the self-assessed energy consciousness and the intention to save electricity ($r = .25$, $p = .028$). All other correlations were not significant.

---

[41] Kesselring, Anne; Pelka, Sabine; Svetec, Erica; Nad, Lucija; Seebauer, Sebastian; Skardelly, Sara; Preuß, Sabine: Slashing the surplus – how prosumers with smart metering respond to regulatory restrictions on self-consumption in Croatia, Conference Proceedings, BEHAVE 2023

Table 4-4: Descriptive statistics from wave 2 and wave 4 about the level of self-consumption for the Croatian pilot

	Survey wave 2 (n = 54)			Survey wave 4 (n = 80)	
	M (SD)	Min, Max (potential range)	Scale interpretation: Higher values indicate …	M (SD)	Min, Max (potential range)
Self-assessed energy consciousness	7.24 (1.32)	5, 9 (1, 9)	higher energy consciousness	7.34 (1.25)	4, 9 (1, 9)
Intention for electricity saving (electricity consumption)	3.56 (1.08)	1, 5 (1, 5)	higher intention to save electricity	3.60 (0.99)	1, 5 (1, 5)
Intention for PV energy use (self-consumption)	3.83 (1.10)	1, 5 (1, 5)	higher intention to use PV energy	3.85 (0.98)	1, 5 (1, 5)
Expected increase in electricity consumption	1.48 (2.44)	-4, 4 (-4, 4)	expectation of higher consumption	0.69 (2.43)	-4, 4 (-4, 4)
Expected increase in PV self-consumption	1.50 (2.15)	-4, 4 (-4, 4)	expectation of higher self-consumption	1.23 (2.30)	-4, 4 (-4, 4)

We asked participants to report their behavior with respect to the policy by implementing four variables (all single items). Specifically, we asked participants about their self-consumption, whether they have turned on additional electrical appliances to buffer PV over-production, and whether they shut down the PV plant to avoid a change of consumer status. For this comparison, we again focus on wave 2 (fall/winter of 2022) and wave 4 (spring of 2023). In wave 4, we also asked whether participants changed their heating system, which may also lead to an increase of electricity consumption (depending on the heating system). The descriptive statistics of these variables for wave 1 and wave 2 are displayed in Table 4-5.

Table 4-5: Descriptive statistics from wave 2 and wave 4 about the response to PV regulation for the Croatian pilot

	Survey wave 2 (n = 54)	Survey wave 4 (n = 80)

Increased self-consumption -2 = decreased a lot, 2 = increased a lot	M (SD) = 0.53 (0.94) n = 49	Min, Max = -2, 2 (potential range = -2, 2)	"I am not sure": 6% (n = 5)	M (SD) = 0.38 (0.97) N = 78	Min, Max = -2, 2 (potential range = -2, 2)	"I am not sure": 2% (n = 2)
Turning on additional electrical appliances	Yes: 61% (n = 33)	No: 26% (n = 17)	Other: 13% (n = 7)	Yes: 63% (n = 50)	No: 23% (n = 18)	Other: 15% (n = 12)
Shutting down the PV plant	Yes: 44% (n = 24)	No: 41% (n = 22)	Other: 15% (n = 8)	Yes: 43% (n = 34)	No: 50% (n = 40)	Other: 8% (n = 6)

During the fall (wave 2), the results indicate no or only a little increase of self-reported PV energy use. This is out of line with the initial objective of the app to nudge self-consumption. By contrast, most participants reported to turn on additional electrical appliances during hours of high PV generation. This serves both a direct financial benefit and the alignment of consumption patterns to the regulatory incentive. The survey also reveals high awareness of the policy. Almost half of the participants considered shutting down their production, and only 15% did not have a clear opinion. In this context, it is noteworthy that the dimensioning of the PV plant during installation is a key determinant on whether participants will be at risk of running a surplus, so it is not surprising that a substantial fraction answered "No". The category *Other* includes the option "I did not think about it" to distinguish. The sample is rather evenly split on whether they consider self-curtailment, which indicates that the policy creates segmentation depending on the households' PV installation and equipment.

The descriptive statistics for survey wave 4 show that participants perceive their self-consumption to be unchanged or increased a little over the first quarter of 2023 (see Table 3-2 in 3.1). Even in the spring, a large proportion of the participants (43%) stated that they considered shutting down the PV plant to avoid the status change. Moreover, the majority (61%) reported having turned on additional appliances to achieve savings despite the over-production of their PV power plant. Similarly, only 28% of participants in wave 4 stated to have *not* changed their heating system. 50% (n = 40) reported that they started to occasionally heat with electricity (air conditioner or electricity heaters), 14% (n = 11) reported using a heat pump since the installation of the PV plant, 5% (n = 4) replaced the gas boiler with an electric one, and 4% (n = 3) chose "Other". These results fit with the other self-curtailment variables and indicate an increase in electricity consumption by most participants after the installation of the PV plant, which points to rebound effects that are indeed incentivized by the policy.

Notably, the responses for increasing consumption and shutting down the plant are on par with those from the fall. Applying a paired t-test as an inference statistical comparison of the self-consumption variable (with n = 48) shows no significant difference. However, the small sample size may limit the comparative analysis. Examining the cross-tables (automatically excluding participants who did not answer the same question in both surveys), it emerges descriptively that only eleven participants (26%) changed their answers across time regarding the shut-down of their PV plant (from yes to no or vice versa, n = 43). The same pattern results for the question to turn on an additional appliance (n = 44): 12 participants (27%) changed their response between survey wave 2 and 4.

4.4 Summary

The statistical evaluation of nudge 1 is not feasible, given the sample size. For nudge 2, we find insignificant effects and discuss the connection to the policy framework in Croatia. The survey data supports the insights from the sensor data: consumers reacted strongly to the policy. The nudging intervention was dominated by policy effects, but the development in Croatia also points to a potential opportunity for nudging. The app allowed consumers to better track their surplus with respect to the policy, so the medium devised originally for the delivery of the nudges served as a medium for monitoring regulatory compliance. Overall, it appears that the NUDGE project provided an additional benefit specific to Croatian users, because it gave transparency and better control to monitor their production and consumption, which is highly relevant for policy. Based on current insights, we suspect this was the driving factor explaining the patterns in the raw data, and we will have to evaluate how the nudges played into this setting.

For nudge 3, we find a positive effect across both groups. However, we acknowledge that the weather is a confounder, especially for group 2 in the regression discontinuity design. The conservative robustness check with the DiD gives a treatment effect of 3.8 percentage points for autarky, which is comparable with the German results (comparison is made to nudges 1 and 2, because nudge 3 in the German pilot is a special case).

5 Belgian pilot: Enhancing knowledge level and decreasing energy consumption

In the following section about the Belgian pilot, two pilot-specific hypotheses are tested. The one on gas consumption is tested based on sensor data, and the one on energy knowledge is based on survey data. The two general hypotheses on intention and motivation are tested based on survey data. As outlined in Table 5-1, we are able to confirm one pilot-specific and two general hypotheses. The section references in the table guide the reader to the analysis based on which the hypothesis is tested.

Table 5-1: Tabular summary for the Belgian pilot

Hypotheses		Based on	Outcome	Section reference
BE1	Energy course positively impact the knowledge level of (a) children attending the energy course and of (b) their parents.	Within-subject with survey data		5.2.2
BE2	The energy course is effective in reducing the gas consumption of the parents.	Within-subject with sensor data (2 or 3 weeks before & after the course)		5.2.6

All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	✓	5.3
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	✓	

5.1 Pilot-specific research design

5.1.1 Hypotheses

The pilot in Belgium is different from the other four pilots because it does not use digital tools like apps or websites. Instead, it implements the NUDGE energy course organized by Spring-Stof, which includes using the EnergyID platform, course materials, and talking with family members at home. The main goal of the Belgian pilot is to see how social influence nudging can help with learning between different generations, with a focus on energy-related issues. Additionally, through use of the EnergyID platform, participants were exposed to feedback and awareness nudges, through access to their consumption data. This assessment is conducted through knowledge tests, relying on self-reported survey data, rather than empirical measurements.

Given this, we formulate:

BE1: The energy course positively impacts the knowledge level of (a) children attending the energy course and of (b) their parents.

Moreover, we looked at the energy consumption of the participating households, in particular the gas consumption, based on consumption data from the EnergyID dashboard. Data was analyzed according to the exact starting date for the lesson.

Given this, we formulate:

BE2: The energy course is effective in reducing the gas consumption of the parents.

We apply the aspect of intergenerational learning also to the general target of increasing the intention and motivation to save energy:

BE3: The energy course positively impacts the intention to save energy within one's household of (a) children attending the energy course and of (b) their parents.

BE4: The energy course positively impacts the motivation to save energy within one's household of (a) children attending the energy course and of (b) their parents.

5.1.2 Analytical strategy

5.1.2.1 Knowledge questions

For the knowledge questions, we used paired sample t-tests, while for energy consumption, we used multilevel regression analysis, with the participants as the random term.

5.1.2.2 Gas consumption

The consumption data was retrieved from the EnergyID dashboard. We paid attention to the intervention periods: the first cohort in February 2022, and the second cohort in December 2022 / January 2023. To understand whether our social nudge had any influence on gas consumption, we selected two timeframes before and after the course took place. We thus perform two separate analyses: a 2-week period and a 3-week period, considering either two or three weeks before and after the exact date of the course. This limited time frame is used to minimize the impact of weather variations.

Figure 5-1: Pre- and post-intervention analysis variations

Moreover, we use two variations of outcome variables: kWh consumption and normalized consumption[42]. For kWh consumption, days with 0 kWh consumed are transformed into missing data, which avoids producing artificially low means. Normalized consumption is calculated by using heating degree days (HDD). This is an often-used approach to account for the fact that people use more energy when it is colder. HDD is calculated by taking a base temperature and subtracting the average daily temperature of that particular day. Base temperatures can vary, with scholars using ranges from 18.3°C[43] up to 23°C[44]. Daily average temperature is calculated by taking the average of the minimum and maximum temperatures. An example calculation of the HDD of dd/mm would be:

---

[42] Ouf, M. M., & Issa, M. H. (2017). Energy consumption analysis of school buildings in Manitoba, Canada. *International Journal of Sustainable Built Environment*, *6*(2), 359-371.

[43] Steemers, K., & Yun, G. Y. (2009). Household energy consumption: a study of the role of occupants. *Building Research & Information*, *37*(5–6), 625–637. https://doi.org/10.1080/09613210903186661

[44] Dombaycı, Ö. A. (2009). Degree-days maps of Turkey for various base temperatures. *Energy*, *34*(11), 1807–1812. https://doi.org/10.1016/j.energy.2009.07.030

HDD = (Base Temperature - Daily Average Temperature)

or

HDD = 23°C - 1°C = 22°C

Using HDD, it is thus also possible to calculate the normalized consumption of a period, which subsequently accounts for the weather. A common application would be to divide the yearly consumption in kWh with the amount of HDDs for the year. Days with HDD lower than 0 (i.e., when it was warmer than the base temperature) equal 0.

In our case, we calculate normalized daily consumption by dividing the consumption for a particular day with the HDD for that day. To prevent losing data in cases where the base temperature is higher than the daily average temperature (i.e.: HDD of 0), we choose a base temperature at the upper end: 23°C, which remains in line with existing work[45], and removes the risk of missing data before and after the intervention periods.

Moreover, both for consumption and normalized consumption, we perform the analysis with and without outliers, given that we have several data points in both cohorts with extremely high values. In this case, the outliers were removed by multiplying the upper interquartile range (IQR) by 5. The IQR is the difference between the 75th and 25th percentiles of the data. IQR was calculated separately for normalized consumption and gas consumption. These data points were then removed from the analysis. We later discuss alternative approaches to minimize the impact of outliers.

To assess the effectiveness of the nudge intervention, we employ random intercept modeling. This approach is appropriate for several reasons. Random intercept modeling enables us to account for individual-level variations that may exist within our study population. By considering these individual differences, we can better understand how the nudge intervention impacts different subgroups or individuals with varying characteristics.

Additionally, random intercept modeling allows us to consider the nested structure of our data, such as repeated measurements or observations within the same individuals. This is particularly important when dealing with longitudinal studies or repeated interventions, as it helps us account for the correlation between observations from the same participant.

5.1.3 KPI & data

For the Belgian pilot, we use two KPIs. The first, knowledge of (a) the pupils and (b) the parents, were captured before the start of the intervention (i.e., the lesson) and again when the course was completed. It consists of a series of knowledge questions. The second, gas consumption, is extracted from EnergyID dashboard and aggregated into daily values.

[45] Dombaycı, Ö. A. (2009). Degree-days maps of Turkey for various base temperatures. *Energy*, *34*(11), 1807–1812. https://doi.org/10.1016/j.energy.2009.07.030

5.1.4 Nudging interventions

As noted earlier, the Belgian pilot does not have a traditional digital intervention but instead relies on intergenerational learning. The rationale of this intervention is that children influence the knowledge of their parents, who will consequently change their behavior[46]. The Belgian pilot consists of two cohorts. These are the courses taking place during the academic years 2021-2022 and 2022-2023. The energy course consists of five lessons taking place over the course of 4-5 months, which were attended by both cohorts. The content of the course covers gas, electricity, and water consumption. Moreover, electricity production and nudging were also covered by the course. During the lessons, pupils are provided with course materials (see nudgeproject.eu) in combination with the EnergyID dashboard. The majority of the pupils are connected to the EnergyID platform, which provides in-depth insights to the energy consumption of one's household.

To determine the start of the intervention (i.e.: defining pre- and post-intervention period) the exact day of the lesson was taken. For our analysis, we consistently apply two variations: assessment of the 14-day pre and 14-day post-intervention and 21 days pre- and 21 days post-intervention, disregarding the exact intervention day (see Figure 5-1). This is done in order to – as much as possible – minimize the impact of the weather, given that in both cohorts, interventions took place during winter and early spring.

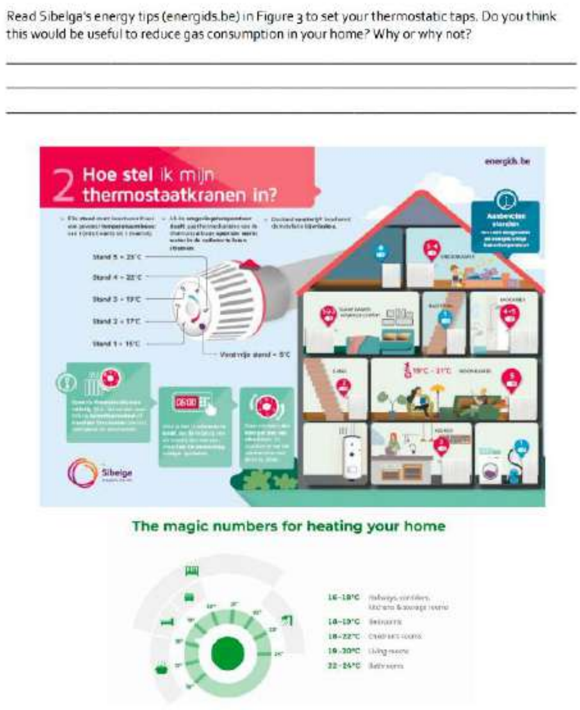


Figure 5-2: An example from the course on energy at SPRING-STOF, with information on correctly setting your thermostat

[46] Damerell, Peter & Howe, Caroline & Milner-Gulland, Eleanor. (2013). Child-orientated environmental education influences adult knowledge and household behavior. Environmental Research Letters. 8. 015016. 10.1088/1748-9326/8/1/015016.

5.2 Testing pilot-specific hypotheses

5.2.1 Descriptive statistics

As noted, the Belgian pilot consists of two cohorts. Cohort 1 contains n=36 of which n=16 households with digitally available gas data, while cohort 2 consists of 40 households, 25 of which submitted gas consumption information. Consumption data can be seen in Figure 5-4, with peaks in winter consumption, especially during the Christmas period, while we also note increases in gas consumption during April. Possible explanations include the start of the easter holiday. Median daily consumption across both cohorts for the entire measurement period is 41.58 kWh, which is lower than mean Flemish consumption of for an average family 64 kWh . However, as illustrated in Table 5-2, consumption during the winter periods is much higher, reaching a median daily consumption of 193.54 kWh in cohort 1 and 91.85 kWh for cohort 2. This is also visible in Figure 5.3, where we see lower mean consumption for cohort 2, compared with cohort 1 (Welch t-test, $t\ (1041.4) = -14.669$, $p < 0.001$). Note also that our data contains several participants with extremely high values. Possible explanations for such high daily consumption might be owning a swimming pool, a sauna, or a large, poorly insulated home that requires more heating during the colder months. Looking at our summary data in Table 5-2, we see that both mean and median consumption is lower for cohort 2, whose intervention period is during the '22/'23 energy crisis.

Table 5-2: Brief description of cohorts in the Spring-Stof pilot

	Cohort 1	Cohort 2
Participants	16	25
Range of dates for the gas lesson	February 7th, 2022 - 18th of February 2022	December 2nd, 2022 - January 9th 2023
Range of data analysis	January 17th 2022 - 11th March 2022	November 11th, 2022 – January 30th 2023
Mean daily consumption	203.50 kWh	117.187 kWh
Median daily consumption	193.53 kWh	91.852 kWh
HDD (base of 23°C)	16.84	17.15

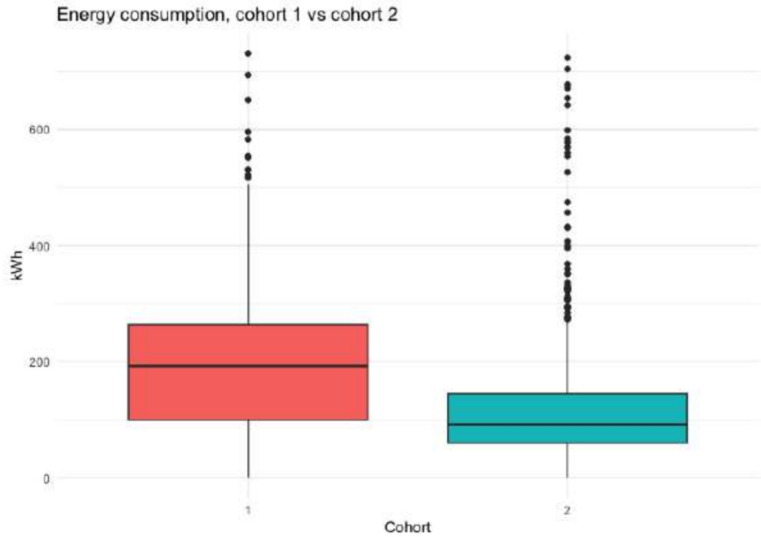


Figure 5-3: Comparing consumption during the six-week measurement period for cohort 1 and cohort 2

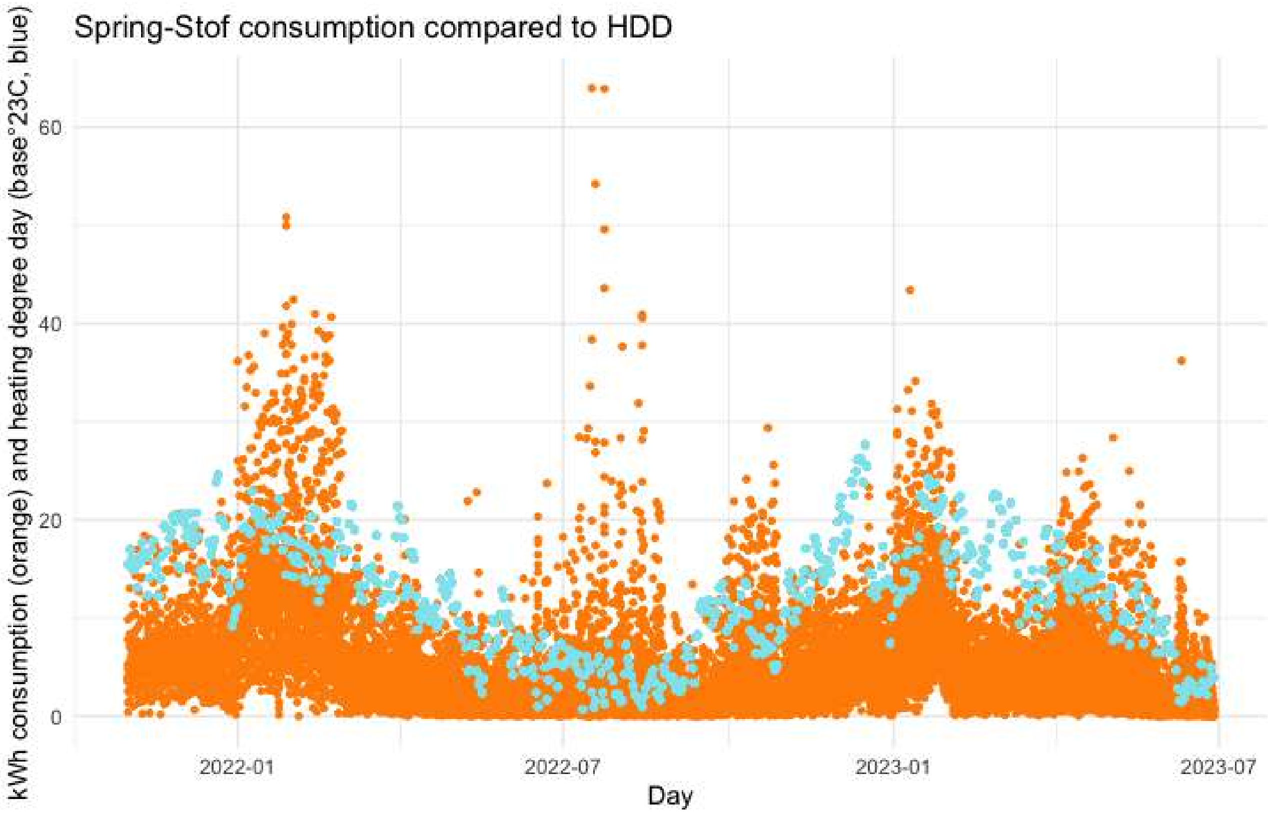


Figure 5-4: kWh consumption (orange) and heating degree days (blue)

5.2.2 Nudge effects

5.2.2.1 Cohort 1 – Two weeks

Table 5-3: Random intercept model results for cohort 1, 2 week variation

	Dependent variable:			
	kWh	kWh (no outliers)	normalized kWh	normalized kWh (no outliers)
	Model 1a	Model 1b	Model 1c	Model 1d
intervention period, 14 days (1)	-52.65**	-47.37**	-2.52**	-0.87**
HDD (23°C base)	-0.68	-0.27		
Constant	242.58**	227.52**	14.07**	8.66**
Observations	409	402	409	220
Log Likelihood	-2,368.15	-2,265.48	-1,269.38	-487.69
Akaike Inf. Crit.	4,746.29	4,540.97	2,546.76	983.39
Bayesian Inf. Crit.	4,766.36	4,560.95	2,562.81	996.96
Notes:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

Our results show that consumption has decreased significantly in the two-week period after the intervention compared to the two-week period preceding the intervention. HDD has no statistically significant impact (model 1a, model 1b), while we see in both our models where outliers were removed that the coefficient decreases (model 1b and model 1d), i.e.: consumption in the post-intervention period was lower than in the pre-intervention period.

5.2.3 Cohort 1 – Three weeks

Table 5-4: Random intercept model results for cohort 1, 3 week variation

	Dependent variable:			
	kWh	kWh (no outliers)	normalized kWh	normalized kWh (no outliers)
	Model 2a	Model 2b	Model 2c	Model 2d

intervention period, 21 days (1)	-72.50**	-68.77**	-3.62**	-1.75**
HDD (23°C base)	-1.54	-0.9		
Constant	266.71**	265.35**	14.11**	8.74**
Observations	608	601	608	332
Log Likelihood	-3,518.18	-3,423.45	-1,872.88	-763.77
Akaike Inf. Crit.	7,046.37	6,856.90	3,753.75	1,535.54
Bayesian Inf. Crit.	7,068.42	6,878.90	3,771.39	1,550.76
Notes:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

Our three-week results are consistent with the two-week analysis: both our kWh consumption and normalized consumption are significantly lower, regardless of whether outliers are included or excluded (models 2a, 2b, 2c, and 2d).

5.2.4 Cohort 2 – Two weeks

Table 5-5: Random intercept model results for cohort 2, 2-week variation

	Dependent variable:			
	kWh	kWh (no outliers)	normalized kWh	normalized kWh (no outliers)
	Model 3a	Model 3b	Model 3c	Model 3d
intervention period, 14 days (1)	40.85**	35.97**	2.18**	2.08**
HDD (23°C base)	6.61**	5.96**		
Constant	-18.88	-7.79	5.71**	5.71**
Observations	626	618	626	624
Log Likelihood	-3,333.12	-3,182.63	-1,531.65	-1,488.31
Akaike Inf. Crit.	6,676.25	6,375.26	3,071.30	2,984.63
Bayesian Inf. Crit.	6,698.44	6,397.39	3,089.06	3,002.37
Notes:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

In contrast with cohort 1, we find in cohort 2 that when considering the two-week intervention period, gas consumption increases, both for kWh consumption and normalized consumption. Outlier removal in both cases have no impact: consumption in the post-intervention period remains higher. Compared to cohort 1, we find that HDD have a statistically significant impact on consumption.

5.2.5 Cohort 2 – Three weeks

We find similar results when moving to a three-week intervention period, with higher consumption in all cases (normalized, with or without outliers). As with our 14-day variation in cohort 2, the inclusion of HDDs has a statistically significant impact on our results.

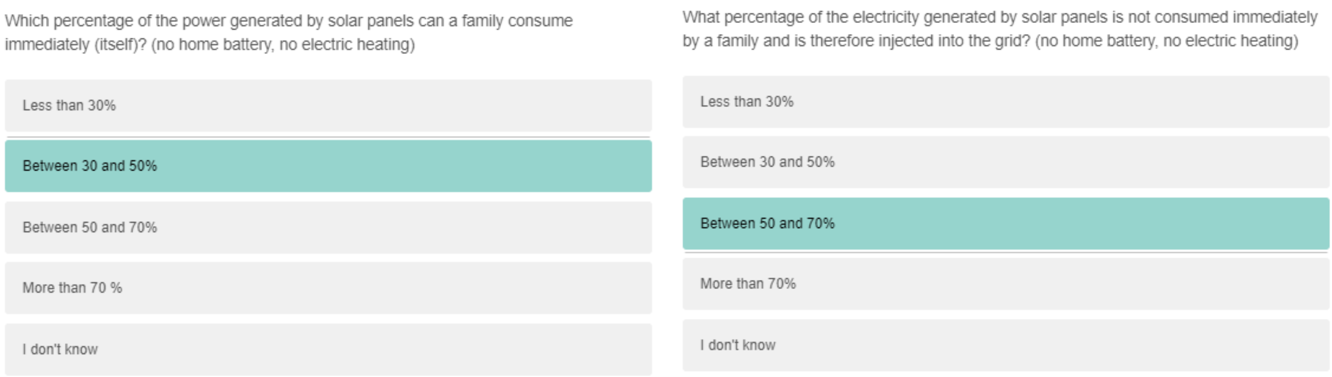
Table 5-6: Random intercept model results for cohort 2, 3-week variation

	Dependent variable:			
	kWh	kWh (no outliers)	normalized kWh	normalized kWh (no outliers)
	Model 4a	Model 4b	Model 4c	Model 4d
intervention period, 21 days (1)	43.88**	38.25**	2.44**	2.36**
HDD (23°C base)	7.27**	6.55**		
Constant	-31.88*	-20.35	5.40**	5.41**
Observations	939	927	939	937
Log Likelihood	-5,041.24	-4,800.91	-2,316.25	-2,269.58
Akaike Inf. Crit.	10,092.49	9,611.82	4,640.49	4,547.16
Bayesian Inf. Crit.	10,116.71	9,635.98	4,659.87	4,566.53
Notes:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

5.2.6 Energy knowledge

To assess the explicit knowledge about energy and water, we have carried out a formal knowledge test consisting of 10 multiple choice questions, which were administered before and after the course period. The topics and content of the questions have been determined based on the course materials and in cooperation with Spring-Stof. To prevent a test-retest effect, we have opted for two equivalent versions of the same test, i.e., the pre-test and post-test. In order to warrant the comparability of both tests, we have implemented equal construct requirements, i.e., the same themes (water & energy, consumption & production) are covered in both knowledge tests, and equal reliability requirements, i.e., content and cognitive processes are balanced, inclusion of the same number of questions and score points, reverse wording of same questions, etc.

In Figure 5-5 an example of the two versions of the same question is illustrated. This question in the post-test returned the least correct answers by pupils (respectively 2.8% and 15% in cohort 1 and 2) and parents (respectively 17.7% and 20.3% in cohorts 1 and 2). The question with the most correct answers was *'A family living in a house built after 2000 consumes the most energy …' (pretest, correct answer: for heating)*, ranging from 59.6% to 61.10% correct answers.



Which percentage of the power generated by solar panels can a family consume immediately (itself)? (no home battery, no electric heating)

Less than 30%

Between 30 and 50%

Between 50 and 70%

More than 70 %

I don't know

What percentage of the electricity generated by solar panels is not consumed immediately by a family and is therefore injected into the grid? (no home battery, no electric heating)

Less than 30%

Between 30 and 50%

Between 50 and 70%

More than 70%

I don't know

Pretest

Posttest

Figure 5-5: Extract from the knowledge pre and posttest, illustrating the comparability of both tests

Repeated-measures ANOVA has been carried out to assess energy knowledge of parents with time as within-subjects measure and condition and cohort as between-subjects variable. We found no effect from time ($F(1,103) = 1.712$, $p = .194$), nor from condition ($F(1,103) = 2.321$, $p = .131$). Cohort significantly impacted energy knowledge among parents with the first cohort scoring significantly higher than the second cohort ($F(1,103) = 5.579$, $p = .020$). A paired samples t-test brought more evidence to this significant difference, which only persists in the pre-test ($t(105) = 2.604$, $p = .011$).

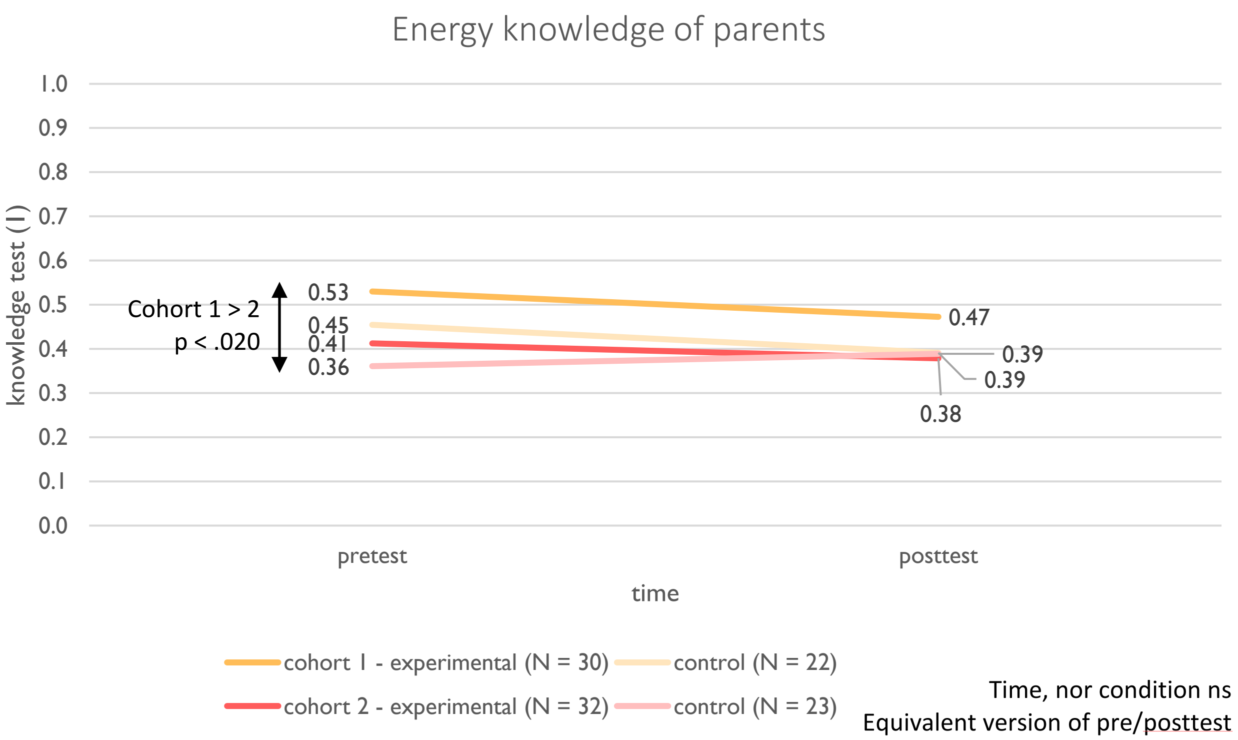


Figure 5-6: Energy knowledge among parents (N = 107)

Also, energy knowledge among pupils has been assessed using a repeated-measures ANOVA with time as within-subjects measure and condition and cohort as between-subjects variable. Neither time, nor cohort have impacted energy knowledge among pupils ($F(1,74)_{time} = .001$, $p = .977$; $F(1,74)_{cohort} = 1.795$, $p = .184$). Pupils' energy knowledge has thus neither increased, nor decreased significantly in response to the energy course.

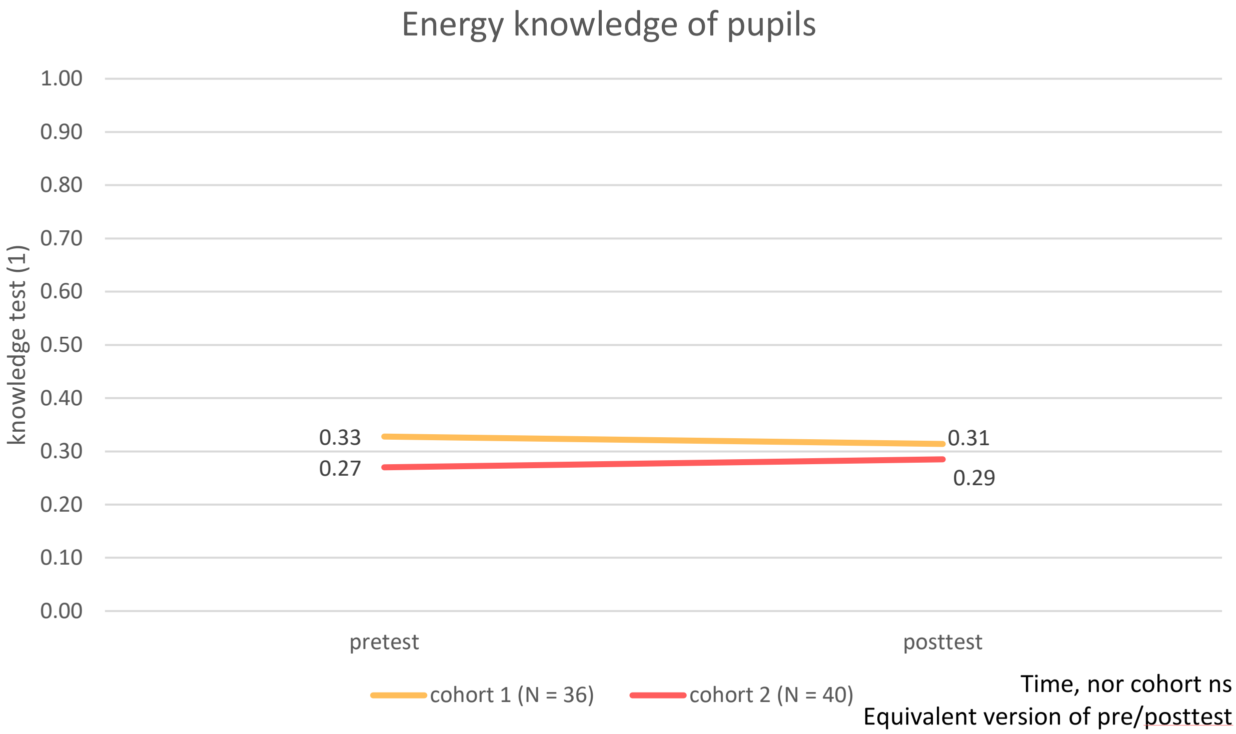


Figure 5-7: Energy knowledge among pupils (treatment group only, N = 76)

Furthermore, we compared the energy knowledge of pupils with their parents in a 2x2 repeated-measures ANOVA with parental relation and time as within-subjects variables and cohort as between-subjects variable. In this pilot no control group with children is included, therefore this analysis only focuses on the treatment group of parents and their children. Parental relation appeared to be a significant predictor of energy knowledge ($F(1,60) = 32.611$, $p = .001$), with parents scoring higher on the knowledge test than their children. This result argues that the potential of intergenerational learning was already limited at the outset of the intervention, since the parent group already scored higher than their child counterparts.

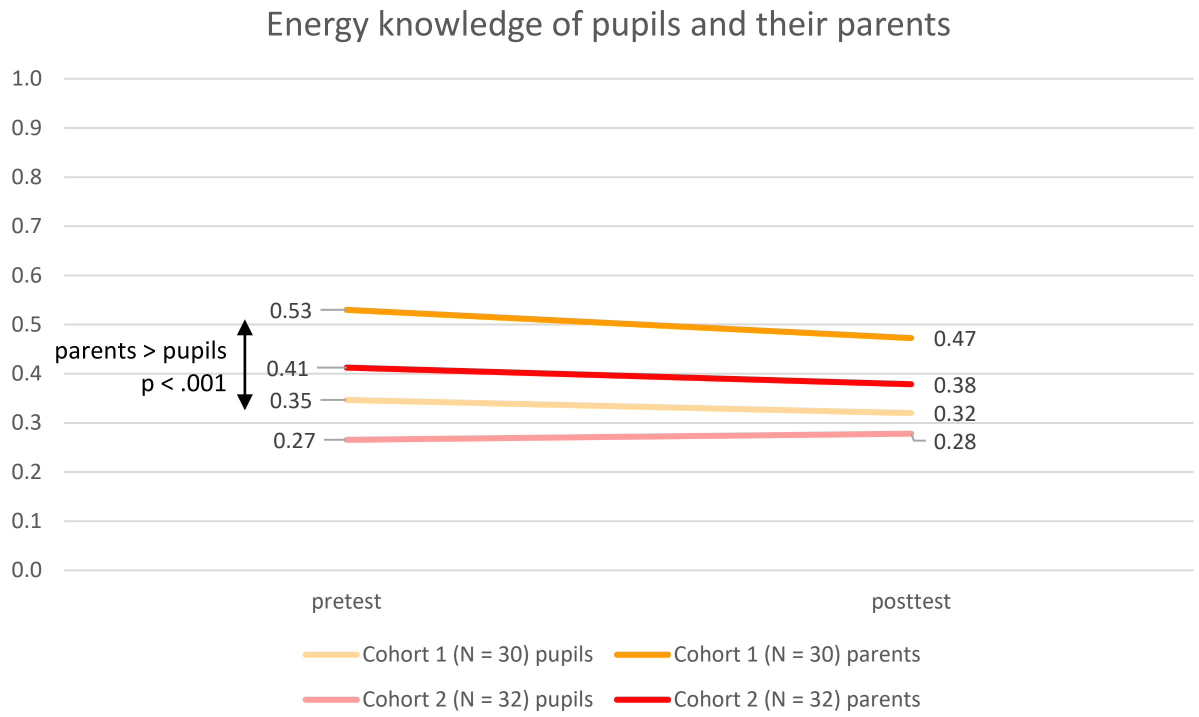


Figure 5-8: Energy knowledge among pupils and their parents (treatment group only)

However, during the second cohort, stakeholders started questioning whether the knowledge test after the intervention could be more difficult than the pre-test. Therefore, we have included an exact retake of the knowledge test in the post-intervention survey of the parents (in addition to the equivalent post-intervention knowledge test). If energy knowledge increases only in the treatment group after the course period, we can argue that the effect can be exclusively attributed to the intervention. Again, repeated measures of ANOVA have been performed with time (within-subjects measure), condition, and cohort (between-subjects measure) included in the analysis. The difference with the analysis presented before is that the pre-test is compared to its exact retake in the post-test. We found a significant increase in energy knowledge in the treatment as well as in the control group ($F(1,52) = 4{,}106$, $p = .048$), indicating there is no significant effect from the condition on energy knowledge ($F(1,52) = 1.821$, $p = .183$). This argues for a test-retest effect since it makes no difference if your child attends the energy course or not.

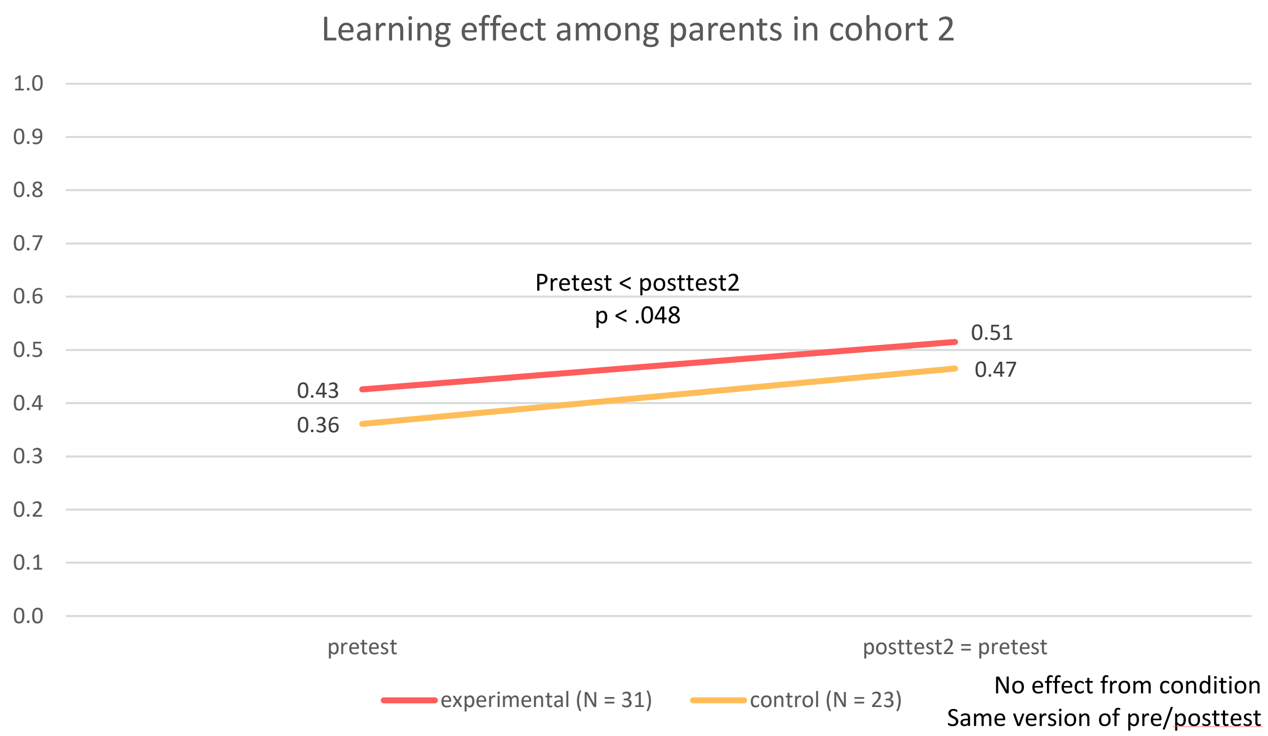


Figure 5-9: Retake of knowledge test in pre- and post-intervention survey

5.3 Testing hypotheses on intention and motivation

5.3.1 Intention to save energy

In both cohorts parents are asked before the intervention to indicate their general intention to save energy and after the intervention to indicate their specific intention to save gas, electricity and water. All intention scores are higher than moderate ranging from 3.27 (post-intervention, intention to save water, cohort 2) to 4.21 (pre-intervention, general intent, cohort 2) on a 5-point Likert scale. Cohort, nor condition have a significant effect on general intent ($t(105)_{condition} = .039$, $p = .969$; $t(105)_{cohort} = .475$, $p = .636$). We have carried out three paired t-tests between on the one hand, general intent and on the other intent to save gas, electricity and water. The intention to save gas, electricity and water is significantly lower than the general intention to save energy ($t(106)_{gas} = 4.041$, $p = .001$; $t(106)_{electricity} = 5.672$, $p = .001$; $t(106)_{water} = 8.226$, $p = .001$). It appears that the intention of parents to save gas, electricity and water drops throughout the course period in comparison to the initial general intention to save energy. However, it should be noted that intent in the pre-intervention survey has been asked on a general level (e.g., *By means of the energy course and the EnergieID dashboard… I intend to save energy*), whereas intention in the post-intervention survey has been asked in relation to a specific energy source (e.g., *I tried to save heating energy at home in the last four months*). Based on these results, it is uncertain to state whether this decrease is due to the specificity of the constructs, and thus the ability of respondents to better assess their specific intention or is a result of the intervention.

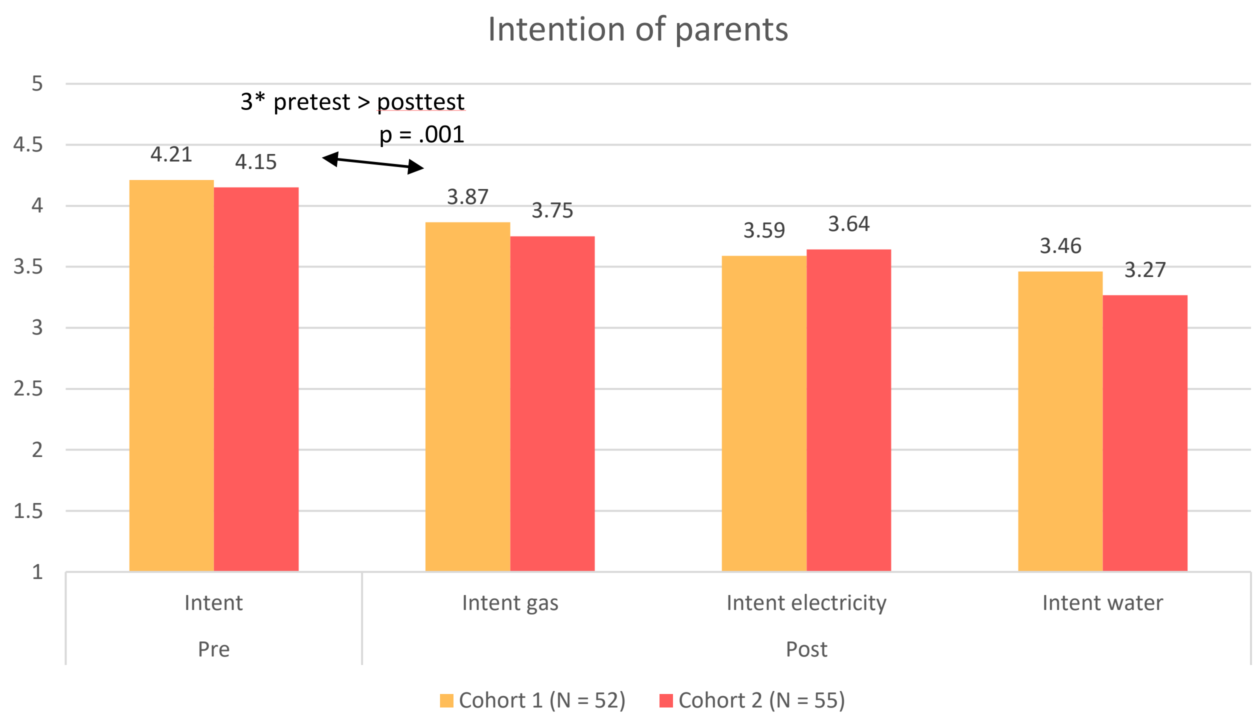


Figure 5-10: Intention to save energy among parents (N = 107)

Intention to save energy has only been assessed in the post-test among pupils. The intention level is moderate, ranging from 3.4 (water, cohort 1) to 3.81 (electricity, cohort 2) on a five-points Likert scale.

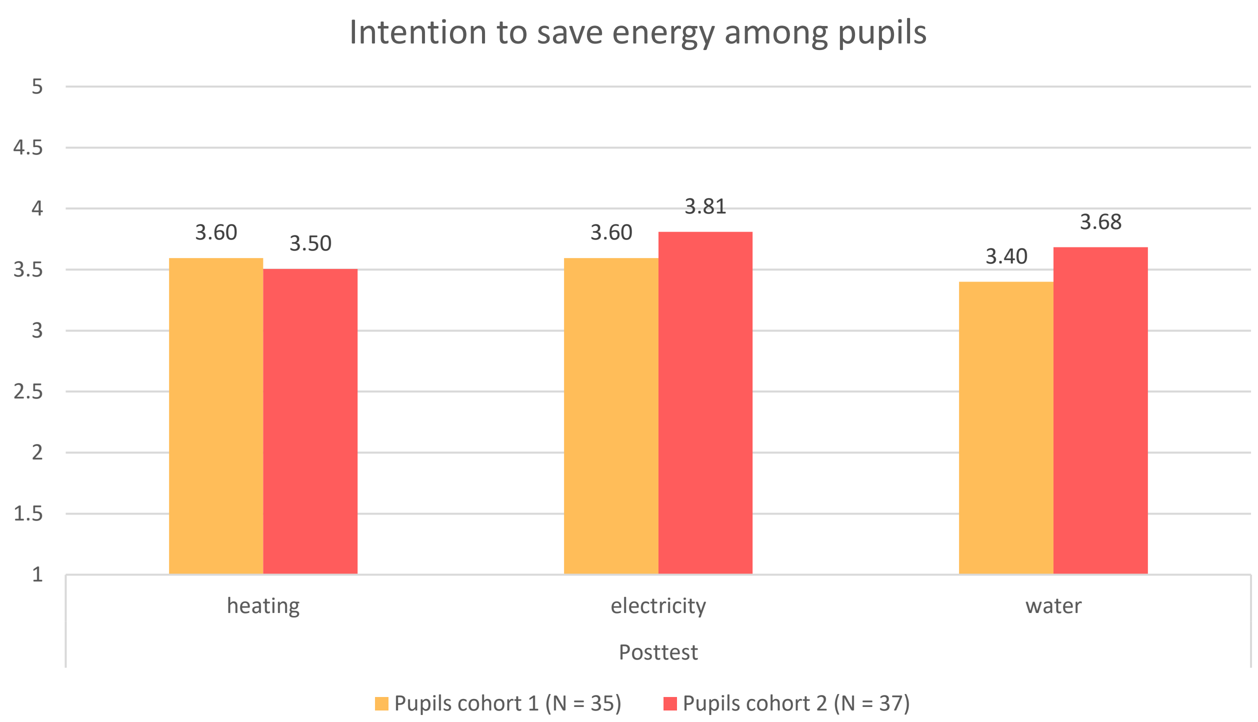


Figure 5-11: Intention to save energy among pupils (N = 72)

5.3.2 Motivation

In both waves, before and after the course period the parents are asked to indicate their intrinsic motivation to save energy (e.g., *I decided to save energy, because … I expected it will be interesting to save energy*). The initial motivation ranges from 2.98 to 3.79 and the resulting motivation ranges from 3.05 to 3.28, which indicates that the motivation scores are further apart before the intervention and move toward each other after the intervention. A repeated-measures ANOVA for motivation has been carried out with cohort and condition as between-subjects factor and time as within-subjects measure. Condition is a significant predictor for motivation to save energy, with the treatment group indicating being more motivated to save energy compared to the control group ($F(1,103) = 5.983$, $p = .016$). The cohort has no significant impact on motivation ($F(1,103) = 2.964$, $p = .088$). A significant effect from time has been found ($F(1,103) = 5.721$, $p = .019$), which suggests that the motivation drops throughout the course period. The interaction effect between time and condition is not significant ($F(1,103) = 2.838$, $p = .095$). Therefore and because of the small sample size, we have carried out paired t-tests per condition and found that the motivation in the treatment groups in both cohorts significantly decreased ($t(61) = 2.880$, $p = .005$), whereas it stayed indifferent in the control groups ($t(44) = .509$, $p = .613$). These results show that after the intervention, the motivation of the treatment groups in both cohorts decreased towards the level of the control groups, which remained unchanged.

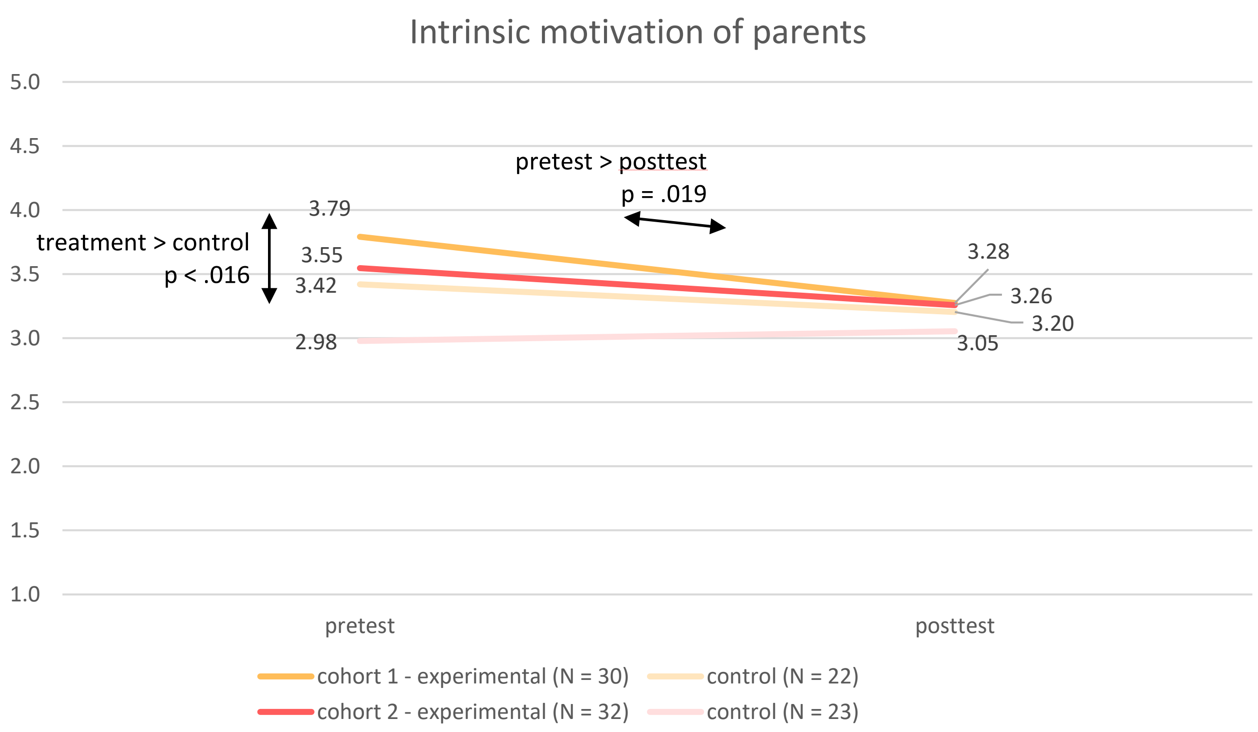


Figure 5-12: Intrinsic motivation among parents (N = 107)

The intrinsic motivation of pupils was assessed before and after the intervention. A repeated-measures ANOVA was administered with time (within-subjects measure) and cohort (between-subjects measure) included in the analysis. We found a significant effect from time ($F(1,73) = 14.542$, $p = .001$). Additional paired-sample T-tests indicated that the effect is not only apparent among the complete sample but also among the second cohort separately ($t(38) = 3.735$, $p = .001$), but not among the first cohort separately $t(35) = 1.676$, $p = .103$). The significant results may be explained by a pronounced strong effect within a smaller sample ($N_{cohort\ 2} = 39$) or a sufficiently strong difference between the pre and post-test within the full sample.

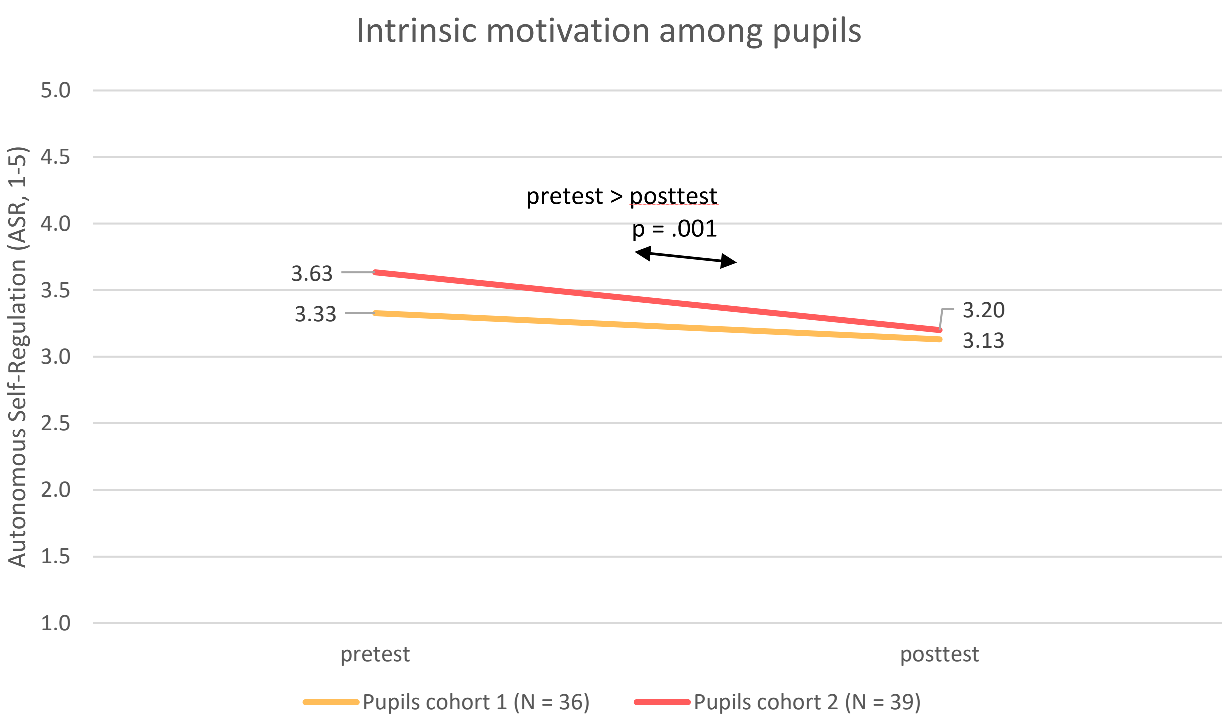


Figure 5-13: Intrinsic motivation among pupils (N = 75)

5.4 Summary

In the Belgian pilot, we have put forward four hypotheses addressing the impact of the energy course on (H1) the knowledge level, (H2) intention to save energy, (H3) motivation to save energy among pupils and parents, and (H4) the gas consumption level of parents. The first three hypotheses have been substantiated by survey data administered before and after the energy course in a pre-and post-test, whereas the fourth hypothesis has been substantiated by gas consumption data before, during and after the intervention period.

Overall, both the results from the self-reported survey data and consumption data are nuanced; none of the hypotheses could be accepted, nor linea recta rejected. The first hypothesis addresses the knowledge level among parents and pupils before and after the energy course. The test scores of parents are significantly lower after the energy course compared to before, the test scores of pupils remained unchanged. However, two possible issues may have caused this surprising result: (1) before the energy course, the test scores of the parents were already higher than those of the pupils, which limited the playground for intergenerational learning from the outset, and (2) the 'equivalent' knowledge testing instruments might have interfered with the testing results, indicating that the knowledge post-test was perceived more difficult than the knowledge pre-test.

The second and third hypotheses dealt with the intention and motivation to save energy. Both have been rejected since parents were significantly less motivated and less intended to save energy after the energy course compared to their motivation and intention before. The motivation of pupils dropped significantly after the energy course.

The fourth hypothesis addresses the impact of the enrolment of pupils in the energy course on the gas consumption in their respective households. Also, these results are somewhat mixed, with a desirable decreased consumption in the post-intervention period for cohort 1 and a higher consumption for cohort 2 in the post-intervention period.

While we find – what appears to be robust – support for our sensor data related research hypothesis in cohort 1, our statistical analysis for cohort 2 raises questions on the reliability of these results. We note that accounting for extreme outliers does not significantly affect our results, indicating that our failure to support our hypothesis for cohort 2 is likely not attributable to extreme outliers found in either cohort 1 or cohort 2.

A few possible explanations might be that the results of cohort 1 are the result of generally higher energy use during the 2021/2022 winter, when compared with the 2022/2023 winter, whereby participants saw more scope for energy reduction. These differences can also be seen in Figure 5-4, where the mean consumption for cohort 2 is significantly lower during the overall measurement period. However, we also note that for cohort 1, the post intervention temperature was higher (i.e.: HDD was lower) and for cohort 2 the temperature was lower (i.e.: HDD was higher).) (see

Table 5-2). While we account for the need to use more energy by 1) controlling for HDD and 2) normalizing consumption, it is possible that in both cases the real-world impact of consumption could not be completely muted.

Another issue that could have had an important impact on our results are the exact starting dates. As illustrated and mentioned in Table 5-2 our starting dates differ for both cohorts. This was due to the necessity of starting the lessons at different times, partly also the result of the large disruptions due to the COVID-19 pandemic. As a result, for cohort 1 the lessons about gas were given between the 7th of February 2022 and the 18th of February 2022 in cohort 1, with the lessons in cohort 2 taking place earlier in the season: the 2nd of December 2022 and the 9th of January 2023. This would subsequently imply that the Christmas season, which is typically a period of increased consumption, would fall in the post-intervention period for cohort 2, while for cohort 1 this would roughly fall the pre-intervention period. This is further compounded by the start of spring in 2023 in the post-intervention period for cohort 2, with corresponding lower HDDs.

6 Greek pilot: Decreasing heat consumption

In the following section about the Greek pilot, two pilot-specific hypotheses are tested based on sensor data. The two general hypotheses on intention and motivation are tested based on survey data. As outlined in Table 6-1, we partially confirm one pilot-specific hypothesis on gas consumption. The other three hypotheses are not confirmed due to data issues. Additional analyses considered the location of the pilots and the exposure of participants to the nudge. However, neither kind of additional analysis changes the overall outcome. The section references in the table guide the reader to the analysis based on which the hypothesis is tested.

Table 6-1: Tabular summary for the Greek pilot

Hypotheses		Based on	Outcome	Section reference
GR1	Nudges are effective in reducing participants' gas consumption.	DiD with sensor data (daily aggregation, nudge 2 vs. nudge 3)	– ✓	6.2.2
GR2	Nudges are effective in reducing the heating time of participants.	DiD with sensor data (daily aggregation, nudge 2 vs. nudge 3)	–	
All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	–	6.3
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	–	6.3

6.1 Pilot-specific research design

Thanks to the generally mild winter in Greece and its Mediterranean climate, the gas consumption is non-zero during winter and late autumn months (see Figure 6-1). This essentially limits the time periods, over which interventions are meaningful, and it was the main reason why the original decision was in favor of a within-subjects design without crossover, i.e., without splitting the households into equal control and treatment groups and switching them in the middle of the intervention period. This design demands half the time for each intervention than its balanced design counterpart, applied in the German, Portuguese and Croatian pilots. Moreover, the number of participating households during the first intervention period was around 40, which means that swapping the control and treatment groups would result in samples of 20 households each, best case.

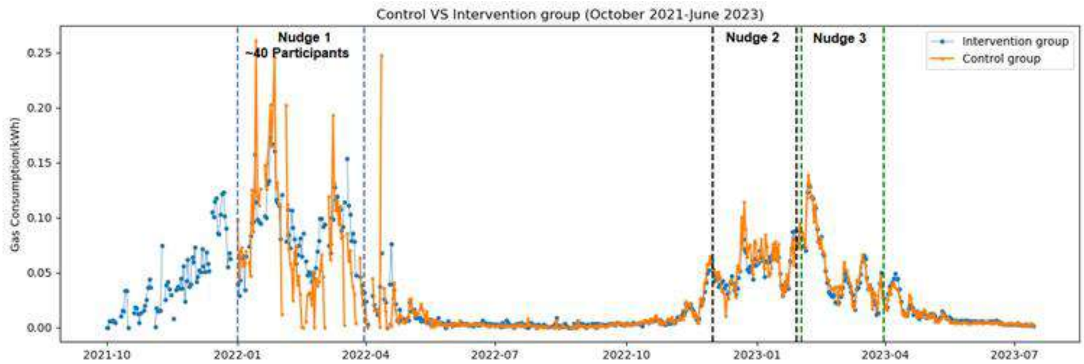


Figure 6-1 : Average gas consumption comparison between control and intervention groups between October 2021-June 2023

However, the completely atypical events, such as the war in Ukraine and its impact on prices, essentially affected the precondition for reliable experiment outcomes and asked for an introduction of a reference case without the nudging interventions. Such a reference would capture the impact over time of the events in winter 2022 and their long repercussions in the energy market and help to disentangle the nudging effect from confounding factors, including weather conditions and price changes.

This reference came in an ad hoc way, rather than after randomized selection, out of the pilot participants who did not use the domX mobile app at all and, hence, were not nudged at all. These participants could be identified thanks to the analysis of data that were logged by the domX mobile app about the interaction of pilot participants with the app and their exposure to screens implementing nudging interventions[47]. Such mobile app data became available during the second and third intervention periods of the Greek pilot thanks to enhancements made to the domX app for this specific reason.

Therefore, with this reference at hand, we could proceed with a solid statistical analysis for the effect of nudge 3. During that third intervention period, both a control group and baseline measurements were available, as required from the Diff-in-Diffs analysis (see section 2.3 and Figure 2-2). In contrast, we could not evaluate the effect of nudge 1 since there was no reference control group for that period in the absence of mobile app logs. Likewise, we could not establish a sufficient baseline for assessing nudge 2 because the gas consumption prior to the launch of nudge 2 (beginning of December 2022) was too low due to the mild weather up to end November in Greece, as shown in Figure 6-1. Therefore, we focus on nudge 3 in the following.

6.1.1 Hypotheses

The nudges in the Greek pilot support the household to reduce their heating consumption. We evaluate the effect on how often the heating is used (heating time) and the amount of gas that is used (gas consumption):

GR1: Nudges are effective in reducing the gas consumption of participants.

GR2: Nudges are effective in reducing the heating time of participants.

6.1.2 KPI & data

For the Greek pilot, there are two main KPIs of interest. First, we focus on the user energy consumption in kWh for heating purposes, *i.e.*, due to the operation of the gas boiler. This is the most relevant variable. It is monitored by specialized sensors in the participating households per minute, and it is summed and reported over intervals of five minutes, as the EnergyIn parameter in Deliverable 3.1, Table 3.

However, the energy consumption is directly reported only for the OpenTherm type boilers (n=83 households) and not for the ON/OFF type boilers (n=19 households). Therefore, we also calculate a KPI that is relevant for all households (n=102). This is the heating *time*, namely the amount of time, during which the

[47] Chitos, Andreas; Karaliopoulos, Merkouris; Pelka, Sabine; Halkidi, Maria; Koutsopoulos, Iordanis: Nudging households for energy savings via smartphone apps: an empirical study, Conference Proceedings, BEHAVE 2023

boiler actively heats the house and energy is being consumed[48]. The heating time is reported every minute, assuming normalized values in the interval [0,1], equal to the percentage of the 1-min interval that the boiler was active[49].

We work with both KPIs at the daily level. Regarding gas consumption, we sum the per 5-min reported EnergyIn values within an hour to compute hourly gas consumption values and then get their daily average. Similar is the process for the heating time, only in this case, the average hourly normalized values result from per minute BoilerHeat variable entries within 1-hr intervals.

As explained in the beginning of Section 6, we had to focus our analysis on the period between late November/early December 2022 to March 2023, namely the time period of the 2nd and 3rd interventions for the Greek pilot, where gas consumption was substantial and the availability of mobile app data made it feasible to extract a control group for the experiment. Therefore, the sensor data analysis is conducted for the specific period, by setting the second intervention cycle as the baseline and the third intervention period as the treatment period, resulting in calculation of the incremental treatment effect of the third nudge.

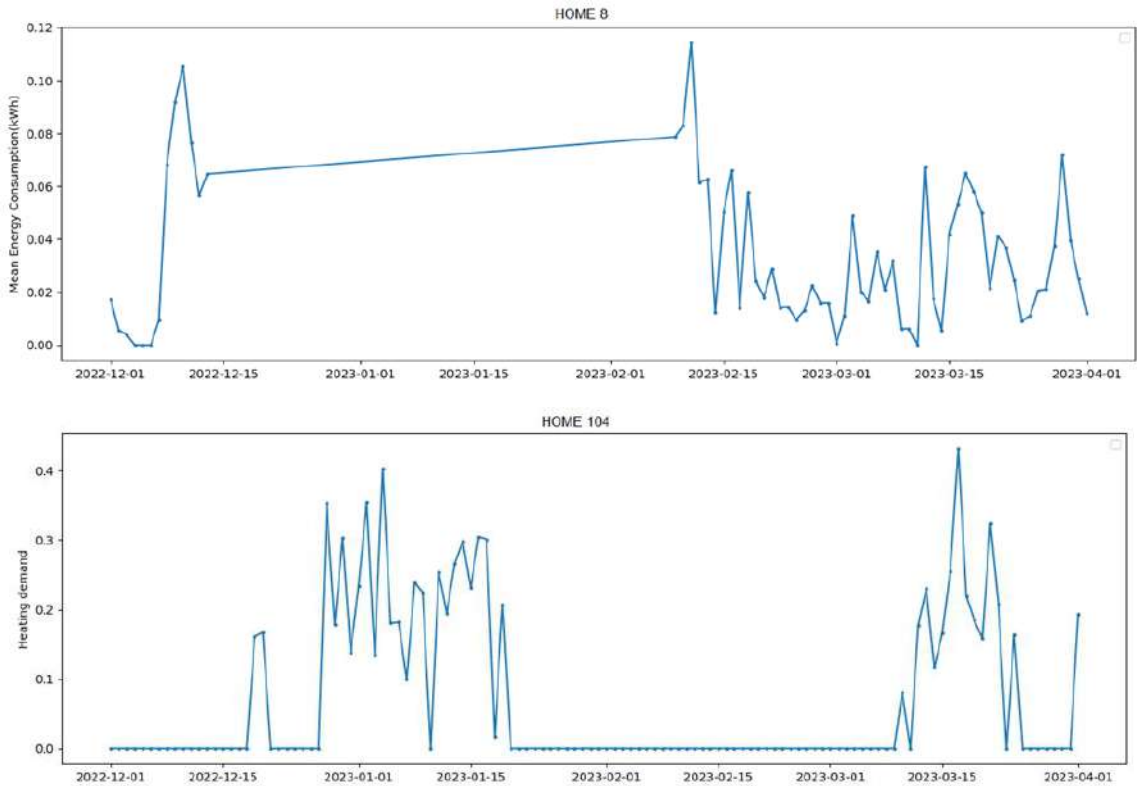


Figure 6-2: Examples of households, which were dropped from the analyzed dataset

---

[48] To convert heating time (in time units) to energy consumption (in energy units), we would need to know the instantaneous value of the modulation of the boiler, but this parameter is not reported for OnOff type boilers (while it is reported for OpenTherm ones).

[49] Please find a more in-depth description of the Boiler heat variable in Deliverables 3.1 and 4.1., https://www.nudgeproject.eu/knowledge-hub/

For several households (n=17) there were long gaps_without receiving measurement data. These households, reported in Table 6-2, were removed from subsequent analysis. All of them were located in Thessaloniki. Moreover, negative-value entries are eliminated from both time series and for the energy consumption time-series, in particular, we also discard entries outside the range [0, 25] kWh, where we normalize the available data with logarithmic transformation. Furthermore, when data are missing for a few days, e.g., for data gaps during the Christmas period, we did not interpolate the missing values.

Table 6-2 : Households participating in the Greek pilot that were left out of subsequent analysis because of full or partial unavailability of their sensor data

Reason for the removal of a household	**Removed households**
No data available between December '22 and March '23	3
More than 20 days without measurement data during the pre-treatment period	3
No measurement data for long period	5
Data available only for January '23	1
No measurement data for February '23	1
Data available from March '23 and on	1
Data available from February '23 and on	3

6.1.3 Analytical strategy

From our analysis of recorded events by the domX mobile app[50], we identify 25 households that have no interaction with the application, and, consequently, they are not exposed to any nudges. Therefore, we identify those 25 households as a potential control group, for the analysis of the pilot's sensor data.

With this a posteriori derivation of a control group (n = 25) against a larger intervention group (n = 75), we could invoke the difference-in-differences (DiD) estimation model with two-way fixed effects (TWFE) to assess the nudging effect while absorbing individual household-specific effects (which remain constant over time) and time-specific effects (which are common across all individual households).

Figure 6-3 and Figure 6-4 suggest that the choice of the control group passes the test of common prior trends. The two groups, control and intervention, trend similarly throughout the Dec '22-Jan '23 period before the 3rd intervention is delivered to them. In the case of energy consumption, we note sharper fluctuations for the control group, which is intuitively expected since its size is 1/3 of the intervention group size.

[50] See Deliverable D1.3 and A. Chitos, M. Karaliopoulos, S. Pelka, M. Halkidi, I. Koutsopoulos, "Nudging households for energy savings via smartphone apps: an empirical study," in Proc. BEHAVE conference, Maastricht, NL, November 2023

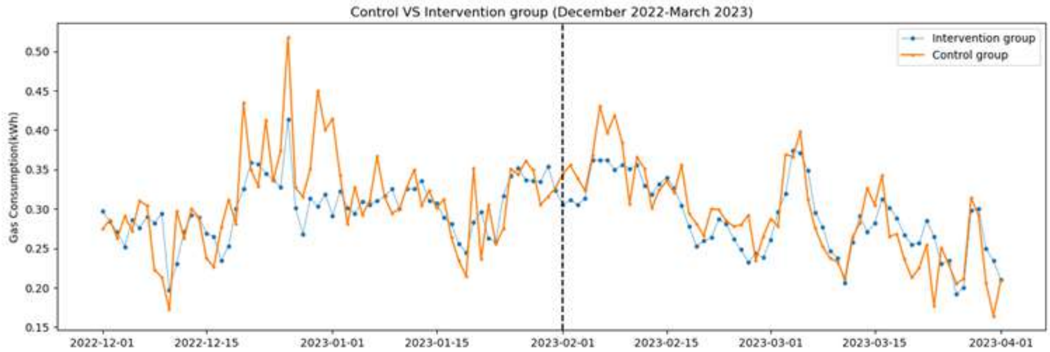


Figure 6-3: Average gas consumption in the control and intervention groups from September '22 till end March '23.

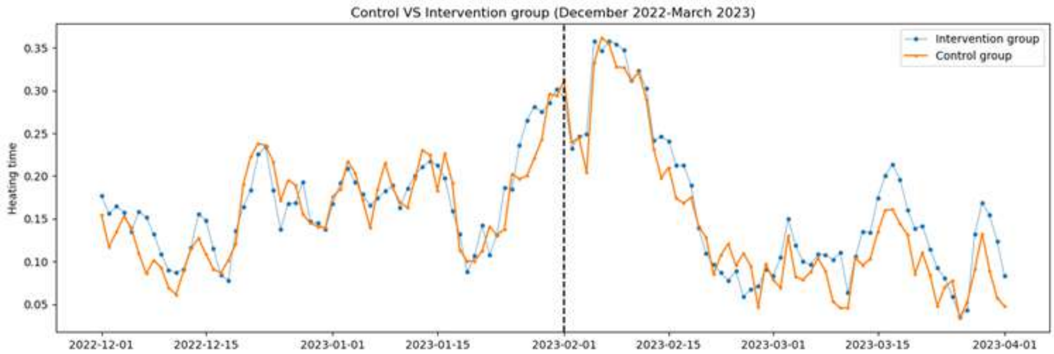


Figure 6-4: Average heating time comparison between control and intervention groups for the second and third intervention periods

Another particularity of the Greek pilot is that the sample is not located in the same geographical area. Table 6-3 reports on the distribution of participant households across 5 different cities. More than 82% of the households though are located in Thessaloniki. Hence, we repeat the DiD analysis, this time limiting the overall sample to households in the city of Thessaloniki.

Table 6-3: Geographical distribution of Greek pilot's participants

Participants' geographical distribution in the Greek pilot					
	Thessaloniki	Athens	Volos	Kalampaka	Karditsa
Participants	83	6	11	1	1

Finally, we experiment with the intervention group specification. Whereas this group nominally involves all 75 households that participated in the experiment, downloaded the app and interacted with it, it exhibits high variance with respect to the frequency of exposure to app features realizing nudging interventions, *i.e.*, how many times the messages and notifications of the 3rd nudge were accessed. Intuitively, users who did not get exposure to such features, were not actually nudged by the app. Hence, for all practical purposes, they cannot be considered part of the intervention group. We have set different thresholds for the number of accesses to the 3rd nudge that marks adequate exposure to it.

Regarding the survey data, we focused on the self-reported intention and motivation to save energy, in line with the analysis carried out for the other pilots in this report.

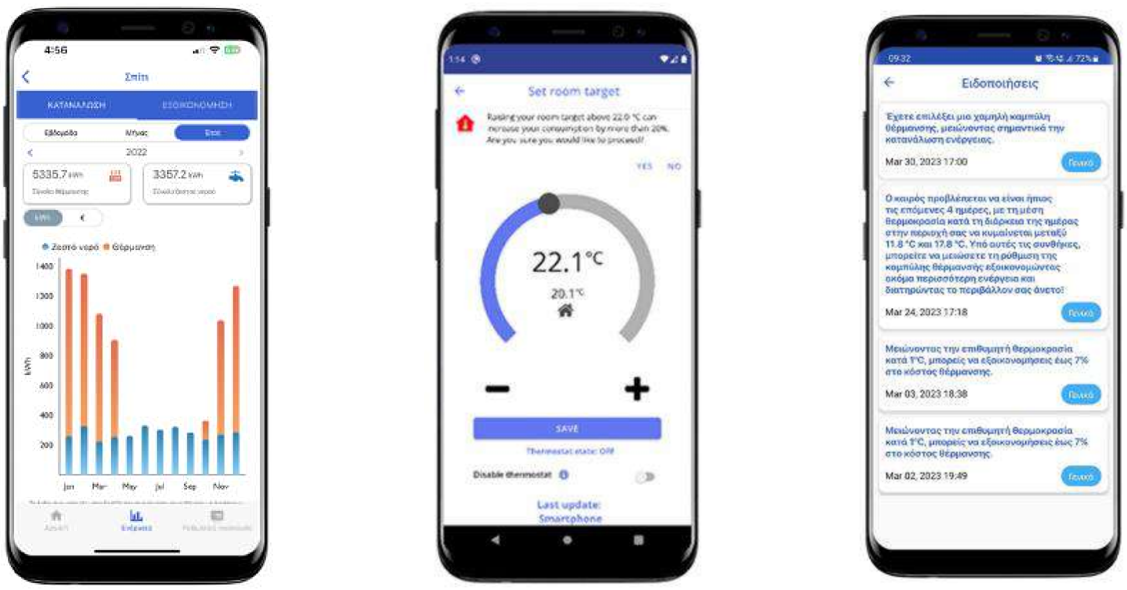
6.1.4 Nudging interventions

Participants are exposed to nudges using the domX application, which is available for mobile devices, and they undergo three nudges, one during each intervention period.

		Nudge 1: Feedback	Nudge 2: Just-in-time prompts	Nudge 3: Push-notifications
Group 1	Baseline	Treated	Treated	Treated
Group 2	Baseline	Control	Control	Control
Timeline	Oct - Dec `21	Jan - Feb `22	Dec `22 - Jan `23	Feb - Mar `23

Figure 6-5: Experiment outline for the Greek pilot

The first nudge is of the feedback and awareness type. The app users are informed about their gas consumption through an app screen featuring a graph bar, energy consumption statistics and comparisons of energy measurements over specific time periods (e.g., day, week, month).



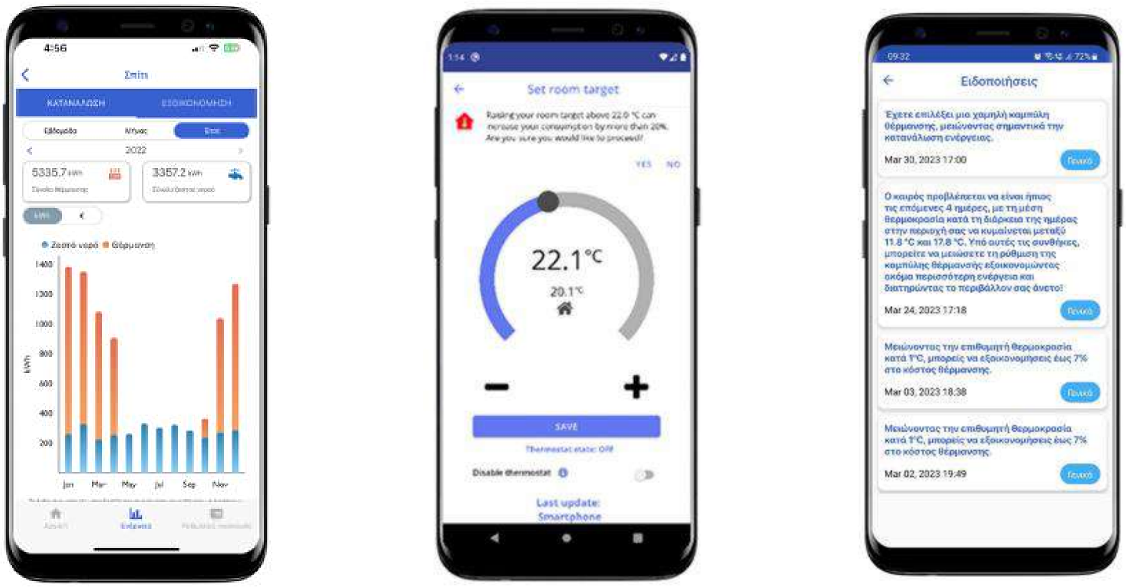
(a) nudge 1 (b) nudge 2 (c) nudge 3

Figure 6-6: Nudging interventions in the Greek pilot

The second intervention is a confrontation nudge, realized through just-in-time prompts. Specifically, when the user tries to either set the room target temperature higher than 21.5° Celsius or disable the weather adaptive-heating feature, an alert box is trying to prevent him/her from accomplishing the action.

The last nudge is realised through push notifications. The notifications are created through messages sent to the devices and stored thereby. In particular, there are two types of notifications, one with an energy-saving tip and another with a congratulating message for properly setting the heating balance. Those notifications and messages were sent three times (3/3/23, 24/3/23, 30/3/23) during the third intervention period.heat

In the Greek pilot, nudges are not removed once they are introduced. This implies that during the second intervention period, pilot participants are exposed simultaneously to both nudges 1 and 2; and during the third intervention period, they are exposed to nudges 1, 2 and 3.



nudge 1 *nudge 2* *nudge 3*

Figure 6-6: shows the domX mobile app screens that implement nudges 1-3.

6.2 Testing pilot-specific hypotheses

6.2.1 Descriptive statistics

Table 6-4 provides basic descriptive statistics for the 2022 heating season, from January '22 to March '22, corresponding to the first intervention period. The intervention and control groups exhibit similar mean values for both KPIs, but the standard deviation is higher for the control group when compared to the treatment group. Additionally, due to the small number of control group's households, we find a significant difference between the two groups regarding the number of available observations.

Table 6-4: Descriptive statistics of daily energy consumption for control and intervention groups in the interval Jan '23 to March '23 (3rd intervention period).

Control group					
KPI	**Mean**	**SD**	**Min**	**Max**	**Observations**
Gas consumption (kWh), n = 16	0.3038	0.2471	7.45x10^-5	1.5581	1835
Heating time, n = 17	0.1688	0.1687	0	1	2259
Intervention group					
KPI	**Mean**	**SD**	**Min**	**Max**	**Observations**

Gas consumption (kWh), n = 61	0.2956	0.1806	9.43x10^-5	1.4446	6773
Heating time, n = 69	0.1802	0.1835	0	1	8851

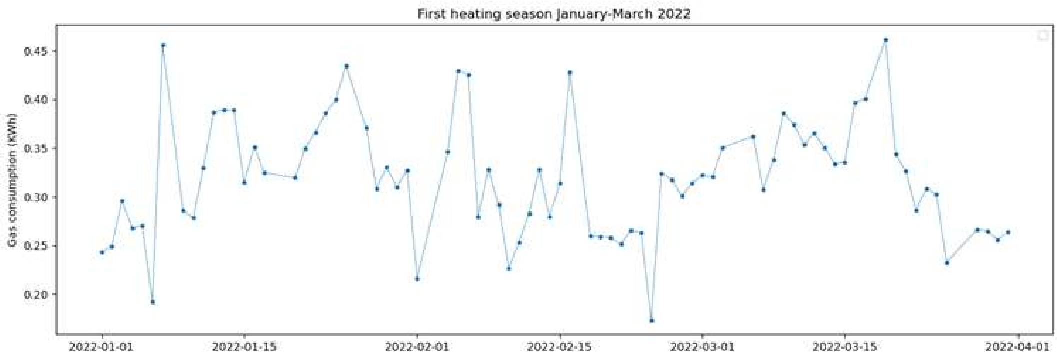


Figure 6-7: Average gas consumption for the first heating season (January-March 2022) during the first intervention period

The average gas consumption is 0.3212KWh. By comparing the available descriptive data for the heating season of 2022 (Table 6-5) with the ones during 2023 (Table 6-4), we note a decrease in the mean values of both KPIs in 2023. This decrease is in line with the 2°C higher mean temperature in the period Jan-Mar' 23 when compared to the same period in 2022 (11°C vs. 13°C in Thessaloniki).

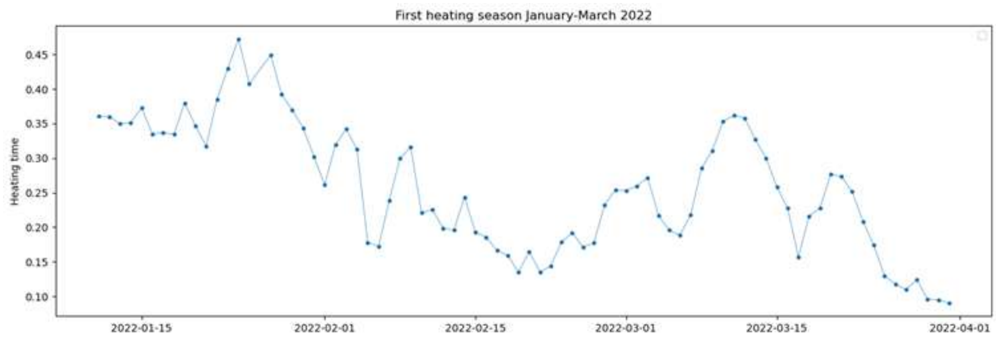


Figure 6-8: Average heating time for the first heating season (January-March 2022) during the first intervention period

Table 6-5: Descriptive statistics of energy consumption at a daily level, as these emerge for the overall sample of 100 households, in the Jan' 22-Mar '22 time interval (1st intervention period)

KPI	Mean	SD	Min	Max	Observations
Gas consumption (kWh)	0.3212	0.2081	6.74x10^-5	1.4037	2104
Heating time	0.2565	0.2013	0	1	2623

6.2.2 Nudge effects

The nudge effect analysis for the Greek pilot is carried out for the period from December 2022 to March 2023, which includes the second and third intervention cycles. The calculated treatment effect relates to nudge 3 so that the model's treatment period coincides with the third intervention cycle (Feb.-Mar.'23) and the interval Dec' 22-Jan '23 serves as the baseline period. During this time, the participating households are subject to the (combined) effect of nudge 1 and nudge 2. From Feb '23 and on, they are exposed to nudge 3 as well.

Table 6-6: DiD results for both KPIs with no effect and fixed effects added to the model for nudge 3

Gas consumption (log(kWh))			
DiD model	**Coefficient of treatment effect, β_3**	**p-value**	**R^2 (between)** [51]
Basic	0.0185	0.697	0.010
+ TWFE	0.0110	0.8621	1.127e-05
Heating time ([0,1])			
DiD model	**Coefficient of treatment effect, β_3**	**p-value**	**R^2**
Basic	0.0128	0.134	0.007
+ TWFE	0.0052	0.6222	8.371e-05

In Table 6-6, we report the estimated coefficients of the treatment effect under the basic DiD models and the variant with two-way fixed effects (TWFE) (see section 9.1.4). In neither of the two cases, do the results imply that nudge 3 had an impact. The estimated effects in both cases are minimal and statistically insignificant. Moreover, the R-squared (R^2) value of the model is remarkably small. The energy behavior of the participants appears to be more affected by the energy prices and the energy crisis in 2022, rather than the notifications realising the nudge 3. Note that the Greek pilot participants persistently point to the increased electricity prices as the main determinant of their energy-saving behavior in all surveys they filled out after interventions, namely wave 2, 3 and 4 surveys (see Deliverable 1.3, section 7.1).

[51] Whereas the DiD models in the previous sections reported the overall R^2 (expressing the explanatory power of the variables including the TWFE), tables in this chapter contains the between R^2 (expressing the explanatory power of the nudge).

6.2.3 Group-dependent effects: Narrowing down the treatment group based on actual exposure to nudging

The pilot participants were exposed to nudges using the domX app. The analysis of the log files that traced their interaction with the app, indicated that approximately 40% of the users was exposed (either pressed a received notification or read a message) 3 times to nudge 3, as it is shown in Figure 6-9. At the same figure, we can see that another approximately 40% of users was not exposed at all to the notifications of nudge 3. Hence, we repeat the analysis of nudging effect reducing the intervention group down to this 40% of users who were indeed nudged through the nudge 3 mechanisms.

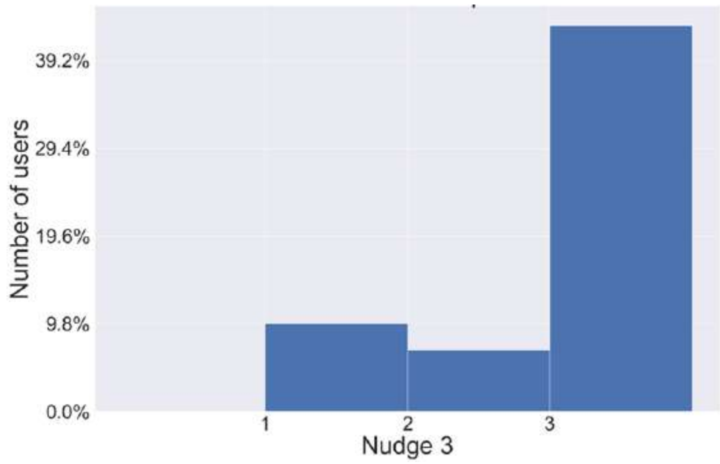


Figure 6-9: Percentage of users exposed to the notifications implementing nudge 3

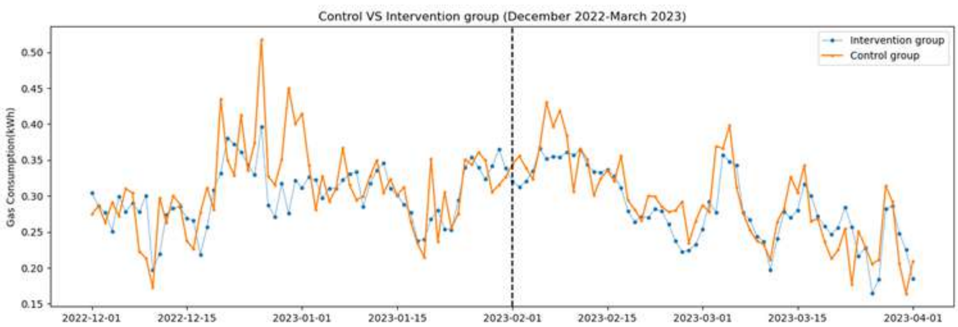


Figure 6-10: Average gas consumption comparison between control and intervention groups, where intervention group consists with users with more than 3 days of nudge exposure

Figure 6-10 plots the aggregate gas consumption of the control group and the reduced intervention group (n=41). The common prior trends requirements is reasonably respected, so we repeat the DiD analysis to assess anew the nudging effect. Table 6-7 and

Table 6-8 summarize the results of this analysis for gas consumption and heating time, respectively.

Table 6-7: DiD analysis for the gas consumption of the reduced intervention group (n=41) after accounting for the exposure of pilot participants to nudge 3

DiD model	Coefficient of treatment effect, β_3	p-value	R^2 (between)
Basic	-0.0314	0.5340	0.0130
+TWFE	-0.0351	0.6215	0.0001

Table 6-8: DiD analysis for the gas consumption of the reduced intervention group (n=41) after accounting for the exposure of pilot participants to nudge 3

DiD model	Coefficient of treatment effect, β_3	p-value	R^2 (between)
Basic	0.0085	0.9440	0.0100
+TWFE	0.0061	0.5707	0.0002

For the analysis provided, we worked with 16 and 33 participants as control and intervention group correspondingly. This time we compute, for first time, a negative coefficient for the gas consumption KPI, which would translate in energy-saving impact on the intervention group, but this is both small and statistically insignificant. For the heating time, the effect is practically zero and not at all significant. Hence, this more careful choice of the intervention group did not alter notably the overall view we have got so far about the (non) impact of nudge 3 on the pilot participants.

6.2.4 Group-dependent effects: households located in a single geographical region (Thessaloniki)

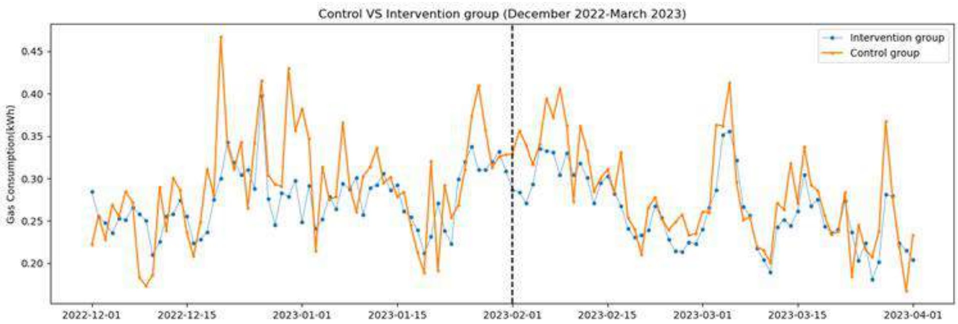


Figure 6-11: Average gas consumption comparison between control and intervention groups for households located in Thessaloniki.

The participant households in the Greek pilot are located in three different regions/cities. The highest portion, approximately 83% of the participants, are located in Thessaloniki, corresponding to 12 households for the control group and 46 households for the intervention when the KPI is gas consumption (OpenTherm-type boilers), and 13 households for the control group and 53 households for the intervention group, when the KPI is the heating time (relevant to both types of boilers, OpenTherm and OnOff).

We repeated the analysis for this subset of households for both KPIs, to eliminate the probability that geographic differentiation factors across the different areas, e.g., weather, might mask the nudging effect.

Looking into the aggregate consumption of the control and intervention groups in this subset of households in Figure 6-11, we can see that they satisfy common prior trends requirement, allowing us to use the DiD technique.

Table 6-9: Nudging effect on gas consumption (logarithmically transformed) for households located in Thessaloniki. DiD analysis carried out with different types of fixed effects.

DiD model	Coefficient of treatment effect, β_3	p-value	R^2 (between)
Basic	0.1605	0.0360	0.009
+ TWFE	0.1742	0.1006	0.0014

The treatment coefficients for gas consumption in Table 6-9 are much higher than those estimated over the full set of households (in Table 6-6), still pointing to effects in the opposite direction and in some cases exhibiting statistical significance. On the downside, the explanatory power of the model remains low. For heating time, the picture is much more similar to what we computed earlier in Table 6-6: namely, small effects that are not significant.

Overall, focusing on the area of Thessaloniki, we cannot find more evidence of nudging effects when compared to the basic results, as we identify a small increase in both gas consumption and heating time of the treatment group.

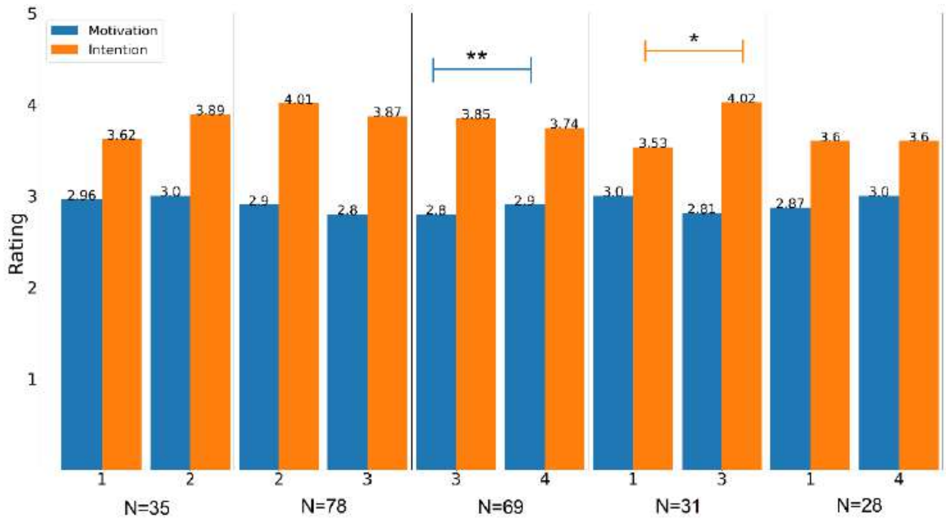
Table 6-10: Nudging effect on heating time for households located in Thessaloniki.

DiD model	Coefficient of treatment effect,	p-value	R^2 (between)
Basic	0.0167	0.088	0.059
+TWFE	0.0116	0.4567	0.0005

6.3 Testing hypotheses on intention and motivation

6.3.1 Intention and motivation

Figure 6-12 compares the mean values of participants' intention and motivation to save energy, as stated in different survey waves. For the first wave, we have fewer answers regarding intention and motivation to save energy, as not all participants are integrated. When comparing the average intention to save energy between the pre- and post-intervention phases, it practically remains the same (Figure 6-12). In between, there is a period that those intentions are boosted, e.g., after the first nudge is applied. However, over the subsequent nudging periods, this impact fades out so that when comparing intentions after all nudges are applied, we can hardly see any difference with respect to the pre-intervention phase.



*Notes: Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01*

Figure 6-12: Mean values of survey responses for intention and motivation to save energy, batched in pairs of survey waves to ease comparisons

Regarding the motivation to save energy, the self-assessment scores of the participants are overall smaller and more stable across the four survey waves. Comparing the pre- with the post-intervention periods, there is a marginal increase of motivation mean value that is not statistically significant.

6.4 Summary

For the Greek pilot, the effects of nudge 1 and nudge 2 could not be analyzed in a reliable manner for different reasons. For nudge 1, we lacked the mobile app data that would allow to separate those users who are exposed to nudges from those who are not and establish a reference control group for running a DiD analysis. For nudge 2, the short winter period in 2022 did not allow to have a baseline period with reasonable gas consumption and the absence of nudges. Analyzing the effect of nudge 3, we could not get a consistently positive view of the nudging effect: the effects were small (and in the opposite direction of the intended one) and statistically insignificant. The significance levels improved and the effect sizes grew, still in the opposite direction, when we narrowed down the analysis on households from the area of Thessaloniki, The effects were in the expected direction only when we filtered the original intervention group according to the level of exposure to nudging, including therein only households with reasonable exposure to the nudge 3 notifications and messages, as it came out from analyzing the log files of the mobile app. Yet, the results still lacked in statistical significance.

The stated intentions and the motivation to save energy remained almost invariable when comparing the self-statements of the pilot participants in the pre-intervention phase with those in the post-intervention period. In light of their persistence with which they name increased energy prices as the dominant determinant of their energy behavior and the overall low levels of interaction with the nudging features of the mobile app, we could not measure a statistically reliable nudging effect in the Greek pilot.

7 Portuguese pilot: Decreasing electricity consumption and improving indoor air quality

In the following section about the Portuguese pilot, one pilot-specific hypothesis is tested based on sensor data. The two general hypotheses on intention and motivation are tested based on survey data. As outlined in Table 7-1, we are not able to confirm the pilot-specific hypothesis but the two general hypotheses. Additional analyses considered the exposure of participants with the nudge and the participants with a thermostat for the effect of nudge 3. However, the additional analyses do not change the overall outcome. The section references in the table guide the reader to the analysis based on which the hypothesis is tested.

Table 7-1: Tabular summary for the Portuguese pilot

Hypotheses		Based on	Outcome	Section reference
PT1	Nudges are effective in reducing participants' electricity consumption.	DiD with sensor data (daily aggregation)	—	7.1.6
All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	✓	7.2
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	✓	7.2

7.1 Pilot-specific research design

7.1.1 Hypotheses

The 101 participants in the Portuguese pilot are equipped with a mobile app (nudge.it) and indoor air quality (IAQ) sensors that aim to increase their awareness of their electricity consumption and IAQ. Each nudge addresses a different KPI. Nudges 1 and 3 aim at reducing heating and electricity consumption, whereas nudge 2 tries to improve the IAQ within the pilot households. The effect of nudging on IAQ and its relation with electricity consumption are analyzed in NUDGE Deliverable 1.3. The KPI we focus hereafter is the electricity consumption making the following hypothesis:**PT1: Households reduce electricity consumption in response to nudging**

7.1.2 KPI & data

The households' electricity consumption is recorded every 30 seconds as the EnergyIn parameter in Wh (as reported in Deliverable D3.1, Table 6). The households participating in the Portuguese pilot are equipped with two different types of power supply: single-phase (n=90) and three-phase (n=11) households. For the three-phase power supply, it is necessary to sum three numbers, i.e., the values reported for each phase to get the household's total electricity consumption every 30 seconds. Furthermore, the electricity measurements are cumulative. Therefore, to calculate the household's hourly electricity consumption, we can subtract the first electricity measurement of each 1-hr interval from the last electricity measurement of the same interval. We then compute the time average of all 24-hourly values to get the average hourly consumption for a single day. The cumulative nature of the measurements reduces the negative impact of missing data since the availability of only two measurements per hour, one close to (ideally: exactly at) the beginning of an hour and another close to (ideally: exactly at) the end of an hour, suffice to compute the hourly consumption.

Additionally, missing and zero energy consumption values were removed from the data, as we normalized the available data with a logarithmic transformation. Out of the available pilot population (n=101), 2 households quitted during the intervention periods, so we worked with sensor data from 99 households, split into two groups, group 0 (n=51) and group 1 (n=48). From those 99 households, 1 household from group 0 is not transmitting electricity consumption data, as it is reported in Table 7-2.

Table 7-2: Portuguese pilot's missing households

Reason for excluding households from analysis	Number of households
Quit the pilot	2
No available electricity data	1

In terms of time periods, the analysis was conducted from the beginning of the pilot's app deployment period (March 2022) until the end of the third intervention cycle (March 2023), without including the washout periods (no intervention) days, as shown in Figure 7-3.

7.1.3 Analytical strategy

To validate our hypothesis for the available population, we worked with the DiD with two-way fixed effects method, which was also used for the analysis of sensor data from the German and Greek pilots. Since the treatment group coincides with group 1 during the first half of each intervention period and with group 0 during the second half of each intervention period, we conduct the analysis separately for each group. As reference period (baseline) for assessing the intervention effect on both groups, we take the pre-intervention period.

As we did with the Greek pilot, we repeated the effect assessment exercise with intervention groups that account for the level of exposure to the nudging features of the mobile app. The level of this exposure was computed by analyzing the data logged by the mobile app about the interaction of end users with the nudging screens of the nudge.it app.

Furthermore, as nudge 3 is aiming to decrease electricity consumption used for heating indoor environment during the cold season, and the available sensor data extends before March 2022 that we considered as baseline period for the other nudges, we repeated the analysis for nudge 3 with an extended baseline period from January to June 2022. The extra added months are considered having high heating demand due to the low outside temperature, so their addition could be valuable for our analysis. However, due to the similarity of those months, we repeat the analysis by using as baseline period only the months of January-March 2022.

7.1.4 Nudging interventions

The Portuguese pilot participants are exposed to nudges through the smartphone application called "nudge.it".

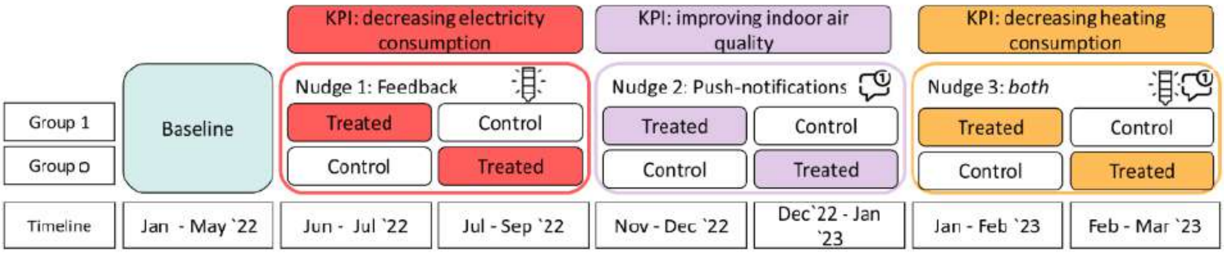
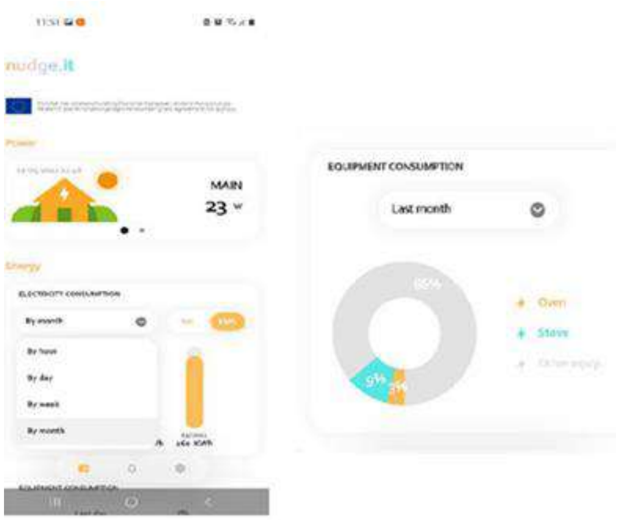
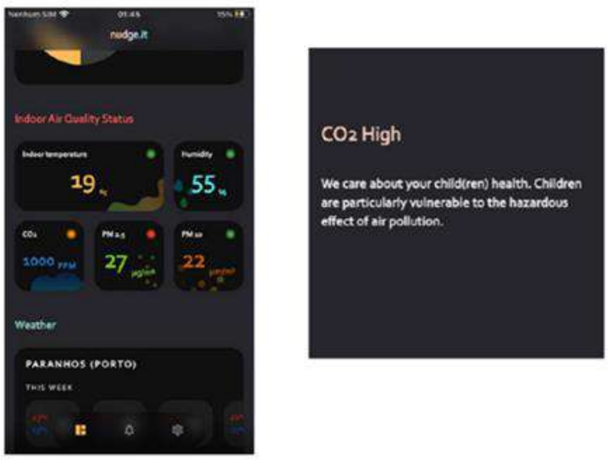


Figure 7-1: Experiment outline for the Portuguese pilot

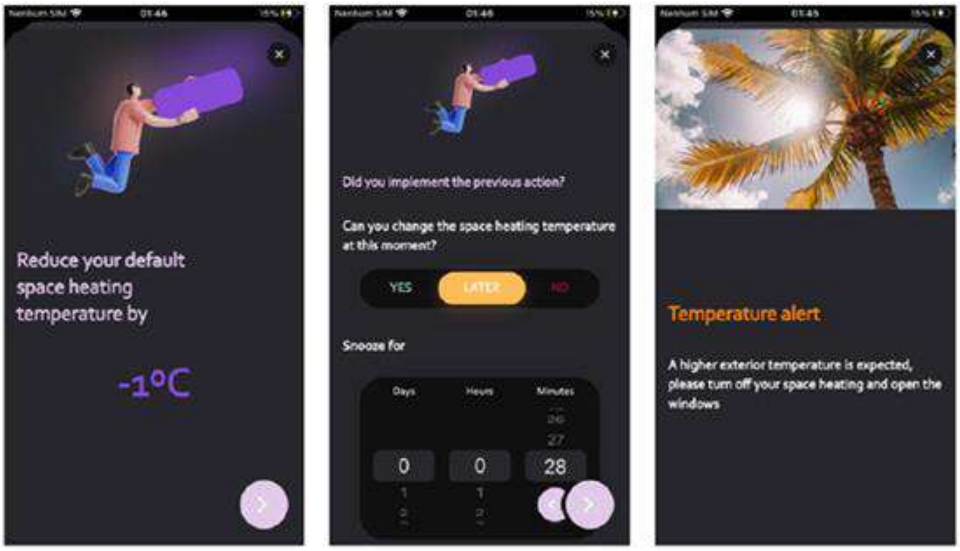
The first intervention is about energy conservation. Users are informed about their electricity consumption over different time intervals through a dashboard, where the electricity consumption can be graphically visualized and compared with different time periods with the use of a pie chart and bar charts. The second intervention targets indoor air quality (IAQ). Measurements of the quantity of (fine) dust particles, PM2.5 and PM10, and carbon dioxide (CO_2) concentration in the air are taken with low-cost sensors and continuously visualized as real-time values in the app. In addition, a push notification is received by users when the CO_2 or PM levels exceed a prespecified exposure limit value (Figure 7-2 (b)).

The third nudge is about adjusting the operational settings of indoor environment heating appliances to reduce electricity consumption through received mobile notifications. However, as the specific nudge is targeting users with a thermostat, a dashboard with real-time data of temperature/ relative humidity and a bar chart indicating the evolution of daily energy consumption during the last 7 days is also available so the users can be informed about their electricity consumption. Moreover, a push notification is received by the users when outdoor temperature is over 2°C than indoor temperature, to turn off the heating system and use the outdoor air as a thermal carrier. Additionally, regarding water and home heating, 77 and 68 households are using electricity and natural gas, respectively. Of those households, 27 have thermostats, and 44 use both electricity and natural gas.





(a) Nudge 1 (b) Nudge 2



(c) Nudge 3

Figure 7-2: Portuguese pilot's app screens for nudging Testing pilot-specific hypotheses

7.1.5 Descriptive statistics

According to the design of the PT pilot experiments, the population of the Portuguese pilot was divided equally into two groups (group 1: n=48, group 0: n=50), so that each group would receive alternately each nudge. Each group is analyzed separately each time as treatment group, to calculate the group's treatment effect per nudge. The comparison of the average daily electricity consumption of both groups (Figure 7-3) indicated that they follow the same energy pattern in the baseline pre-intervention period (common prior trends), allowing us to calculate the treatment effect with the use of the DiD technique.

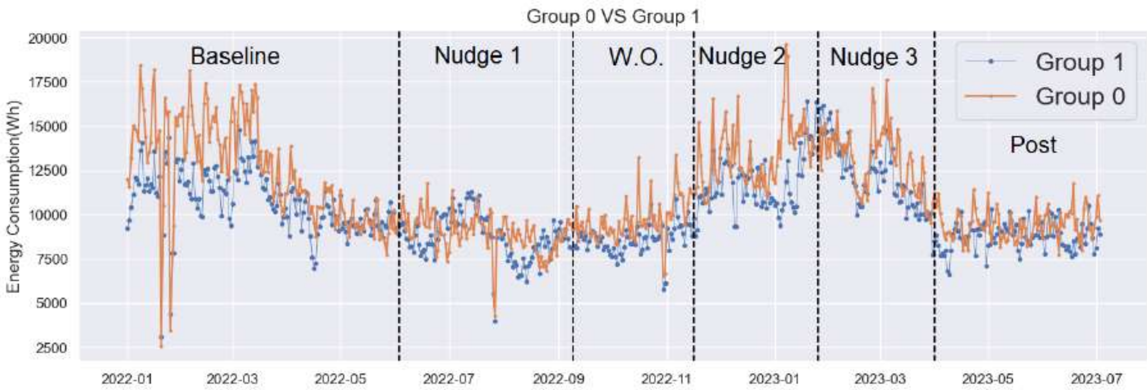


Figure 7-3: Average daily energy consumption for group 1 and group 0

The descriptive statistics of the two groups indicate that group 0 consumes slightly more electricity than group 1, in terms of average daily electricity consumption. Additionally, the electricity consumption of group 1 presents a higher variance than group 0.

Table 7-3:Descriptive statistics of average daily electricity consumption (in Wh) for the two groups in the Portuguese pilot

Group	Mean	SD	Min	Max	Observations
Group 1	11111.02	10185.96	4.8	107212.1	11989
Group 0	12161.06	9271.33	1.2	84720.7	11513

7.1.6 Nudge effects

The conducted analysis covers a period of one-year, from March 2022 (pre-intervention period) to March 2023 (end of third intervention period). The underlying regression model is in described in section 9.1.5.

Table 7-4 reports the coefficient β_3 of the interaction term (intervention group membership times intervention period in 9.1.5), as carried out for the two groups. Overall, the DiD model produces a small R^2 number. Moreover, most of the available results are statistically insignificant, as we get p-values over 0.05.

Table 7-4: Coefficients of treatment effect for the basic DiD model and its two-way fixed effect variation for electricity consumption (in log(Wh)); R^2 is presented as R^2 (between)

Group 1 (n=48)			
DiD model	**Nudge 1**	**Nudge 2**	**Nudge 3**
Basic (log(Wh))	0.0458 (p=0.141, R^2=0.009)	-0.1510 (p=-0.0000, R^2=0.026)	-0.1382 (p=0.000, R^2=0.036)
+TWFE (log (Wh))	0.0692 (p=0.1440, R^2=0.0009)	-0.1258 (p=0.0471, R^2=0.008)	-0.1016 (p=0.1618, R^2=0.002)
Group 0 (n=50)			
	Nudge 1	**Nudge 2**	**Nudge 3**
Basic (log (Wh))	-0.0328 (p=0.233, R^2=0.031)	0.1286 (p=0.000, R^2=0.04)	0.0682 (p=0.025, R^2=0.023)
+TWFE (log (Wh))	-0.0222 (p=0.6655, R^2=0.0001)	0.0851 (p=0.2044, R^2=0.0019)	0.0525 (p=0.397, R^2=0.0008)

When focusing on the two-way fixed effect model in Table 7-4, we note opposite patterns regarding the treatment effects on group 1 and group 0. Hence, nudges 2 and 3 yield positive effects that are both non-negligible and statistically significant for group 1, i.e., the group that is first exposed to each nudge, whereas the effect is in the opposite direction for group 0. This implies learning effects in action, namely the effect of nudges 2 and 3 on the group that is subject to nudging in the first half of the intervention period appear to be lasting over the second half of the intervention period.

In contrast, the effects of nudge 1 are insignificant for both groups. Recall that the Portuguese participants, as the case was with the Greek pilot as well, identify the increased energy prices as the no 1 determinant of their energy-saving behavior throughout the wave 2, 3 and 4 surveys.

In addition to the analysis with the basic DiD model described in the previous section, we considered extending the DiD model with household and time fixed effects, adding dummy variables for individual households and days covered by the intervention periods to the DiD equations in section 9.1.4. Table 7-5 reports the coefficients of nudging effects for the two cases. With time fixed effects, the coefficients of nudging effects are similar in value with those estimated with the basic DiD model (no fixed effects) in Table 7-4, with better significance scores. The way these effects are distributed over the two groups resembles what we discussed earlier for the basic DiD model. In contrast, the addition of household fixed effects yields effects of similar sign but statistically insignificant.

Table 7-5: Coefficients of treatment effect for DiD + time and household fixed effects for electricity consumption (in log (Wh)).

Group 1			
	Nudge 1	**Nudge 2**	**Nudge 3**
+ household FE	0.0681 (p=0.1607, R^2=0.0085)	-0.1281 (p=0.0450, R^2=0.0278)	-0.1060 (p=0.2236, R^2=0.0628)
+ time FE	0.0492 (p=0.0288, R^2=0.0067)	-0.1488 (p=0.0000, R^2=0.0150)	-0.1344 (p=0.0000, R^2=0.0133)
Group 0			
+ household FE	-0.0195 (p=0.7015, R^2=0.0545)	0.0905 (p=0.1802, R^2=0.0689)	0.0588 (p=0.3420 R2=0.0320)
+ time FE	-0.0357 (p=0.0507, R^2=0.0065)	0.1241 (p=0.0000, R^2=0.0142)	0.0627 (p=0.0005, R^2=0.0110)

7.1.7 Deep dive: nudge 3 effect with different baseline period

We repeated the analysis of nudge 3 impact by extending the baseline period from January to May 2022. As the third intervention is targeting the electricity consumption through heating, the added months of January and February will contain valuable data for electricity consumption due to the outside temperatures and the increased heating demand in the extended period.

Table 7-6: Coefficients of treatment effect of nudge 3 for group 1 in the PT pilot with an extended baseline period. DiD model with/without fixed effect, dependent variable is electricity consumption (in log (Wh))

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	**-0.1032**	0.0000	0.0220
+ household FE	-0.0719	0.3533	0.0238
+ time FE	**-0.0922**	0.0000	0.0151
+ TWFE	-0.0569	0.4589	0.0006

From the provided results, we identify a similar pattern with the basic analysis, where group 1 experiences a positive and group 0 a negative effect on its electricity consumption. These effects are significant under the basic DiD model (without fixed effects) and the DiD + time fixed effects, as before.

Table 7-7: Coefficients of treatment effect of nudge 3 for group 0 in the PT pilot with an extended baseline period. DiD model with/without fixed effect, dependent variable is electricity consumption (in log (Wh))

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	0.0188	0.5410	0.014
+ household FE	0.0074	0.8982	0.0027
+ time FE	0.0036	0.8387	0.0155
+ TWFE	-0.0186	0.7443	0.0001

For group 0, on the other hand, all effects are render insignificant when considering the extended baseline period, as shown in Table 7-7.

Focusing further the baseline period on the January-March 2022 interval to carry out a one-to-one comparison between the same months in 2022 and 2023. For group 1, we get a small increase in the size of the nudging effect when we include both entity and time effects, but no statistically significant results for all the available models.

Table 7-8: Coefficients of treatment effect of nudge 3 for group 1 in the PT pilot with baseline period taken to be Jan-Mar 2022. DiD model with/without fixed effect, dependent variable is electricity consumption (in log (Wh))

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	-0.0097	0.7480	0.0260
+ household FE	0.0272	0.7219	0.0036
+ time FE	-0.0074	0.7037	0.0256
+ TWFE	0.0279	0.7129	0.0002

On the other hand, for group 0 we get a positive effect, contrary to the previous analysis steps. However, these effects are only significant under the model with the time fixed effects.

Table 7-9: Coefficients of treatment effect of nudge 3 for group 0 in the PT pilot with baseline period taken to be Jan-Mar 2022. DiD model with/without fixed effect, dependent variable is electricity consumption (in log (Wh))

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	-0.0518	0.1400	0.020
+ household FE	-0.0984	0.1623	0.0031
+ time FE	-0.0549	0.0045	0.0214
+ TWFE	-0.0964	0.1733	0.0020

7.1.8 Deep dive: nudge 3 effect for users with thermostat in their households

Nudge 3 is aiming to reduce electricity consumption by reducing heating consumption. Therefore, we try to repeat our analysis for nudge 3, with baseline period from March-June 2022, focusing on the households that utilize thermostats for setting heating temperature. From the pilot's population, only 23 households are using a thermostat, which are almost equally divided into the two groups (group 1 n=10 users, group 0 n=13 users).

Table 7-10: Coefficients of treatment effect of nudge 3 on households with thermostat for group 1 in the PT pilot. DiD model with/without fixed effect, the dependent variable is electricity consumption (in log (Wh))

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	-0.1935	0.0000	0.0570
+ household FE	-0.1353	0.4825	0.0822
+ time FE	-0.1875	0.0000	0.0113
+ TWFE	-0.1274	0.5067	0.0045

Table 7-11: Coefficients of treatment effect of nudge 3 on households with thermostat for group 0 in the PT pilot. DiD model with/without fixed effect, the dependent variable is electricity consumption (in log (Wh)).

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic (no FE)	0.2668	0.0000	0.039
+ household FE	0.1936	0.1791	0.0587
+ time FE	0.2601	0.0000	0.0130
+ TWFE	0.1852	0.2052	0.0108

From the results in Table 7-10 and

Table 7-11, we observe the same pattern that we identified for the basic analysis. Furthermore, for the basic (no effect) and time effect models where statistically significant results occur for both groups, we can pinpoint an increased effect, either positive or negative, compared to the basic analysis for the third intervention period.

7.1.9 Group-dependent effects: Treatment group based on the days of exposure to nudging

Similar to the analysis conducted for the Greek pilot in Section 6.2, as the Portuguese pilot participants are also exposed to nudges through a mobile application, we asked ourselves how many days of exposure could serve as a threshold *thr* marking adequate exposure to nudging. We tested *thr* values of 1 day, 2 days, 3 days and 4 days for nudge exposure and we report the analysis for *thr*=3 days, which results in an intervention group size that is the average of that for *thr* =1 and *thr* =4 days.

Table 7-12: Coefficients of treatment effect for group 1 in the PT pilot. DiD analysis/without fixed effects and TWFE, the dependent variable is electricity consumption (in log (Wh)). Group 1 includes households with more than 3 days of exposure to nudging.

Model	Nudge 1	Nudge 2	Nudge 3
Basic (no FE)	0.1149 (p=0.027 R2=0.005)	-0.1316 (p=0.0010 R2=0.024)	-0.3424 (p=0.000 R2=0.038)

+ TWFE	0.1039 (p=0.0771 R2=0.0009)	-0.1050 (p=0.3390 R2=0.0016)	-0.2005 (p=0.1807 R2=0.0048)

Compared to the basic results in the previous section, the modified intervention group 1 has an even smaller increase in consumption nudge effect, but with no statistically significant results in most cases. Particularly, during nudge 2, group 1 has a very small decrease in energy consumption treatment effect for all the models. On the other hand, group 0 appears to have a small positive increase compared to the basis analysis we conducted. Therefore, reshaping the treatment groups appears not to be very effective in terms of treatment effect.

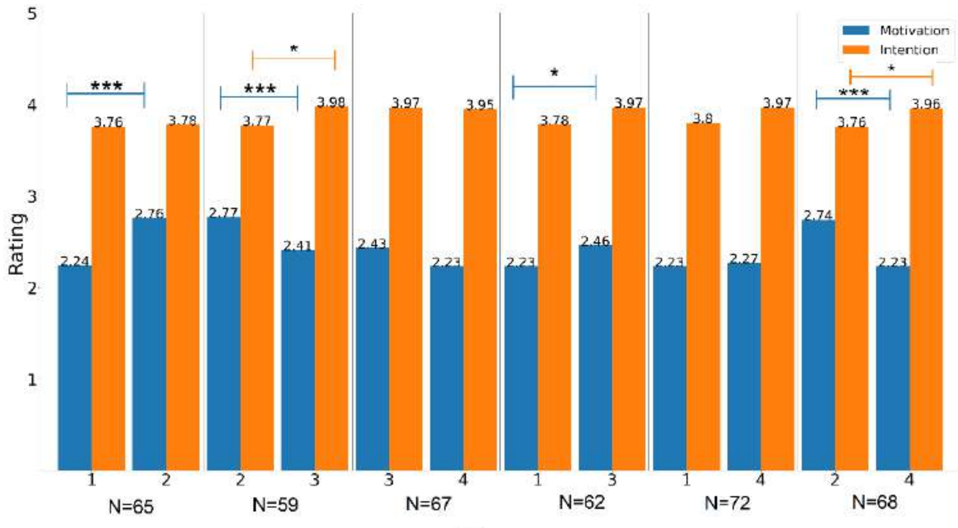
Table 7-13: Coefficients of treatment effect for group 0 in the PT pilot. DiD analysis/without fixed effects and TWFE, the dependent variable is electricity consumption (in log (Wh)). Group 0 includes households with more than 3 days of exposure to nudging.

Model	**Nudge 1** (Obs = 11020)	**Nudge 2** (Obs = 8220)	**Nudge 3** (Obs = 6734)
Basic (no FE)	-0.0359 (p=0.409, R^2=0.030)	0.1567 (p=0.0000, R^2=0.035)	-0.0056 (p=0.911, R^2=0.016)
+ TWFE	-0.0271 (p=0.7638, R^2=0.0001)	0.1126 (p=0.1586, R^2=0.0037)	0.0046 (p=0.9601, R^2=0.0001)

7.2 Testing hypotheses on intention and motivation

For the Portuguese pilot, the mean value of intention to save energy increased from the pre-intervention survey to the post-intervention period (Figure 7-4). The increase was small but significant at the 10% level and it is happening as a whole during the 2nd intervention period, i.e., between the 2nd and 3rd survey wave. There seems to be no change in the 3rd intervention period, that is the pilot participants state make almost identical statements about their intention to save energy in the 3rd and 4th survey waves.

The motivation to save energy scores much lower, not only with respect to the intention scores in the PT pilot but also when compared with the motivation levels stated in the other pilots. When comparing the motivation statements between the pre-intervention and the post-intervention periods, we can hardly see any difference. However, the stated motivation makes a statistically significant jump of 0.5 points during the 1st intervention period, but these points gradually get lost during the 2nd (mainly) and the 3rd intervention period. We note that the PT pilot participants, much as the GR and BE pilot participants did, named the increased energy prices in 2022 as the main determinant of their energy-saving behavior throughput the pilot experiments (all three intervention periods). Hence, we expect that the trends in the intention and motivation scores rather reflect how their concerns about energy prices have fluctuated over time rather than the impact of the nudging interventions.



*Notes: Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01.*

Figure 7-4: Mean values of survey responses for intention and motivation to save energy, batched in pairs of survey waves to ease comparisons: PT pilot.

7.3 Summary

For the Portuguese pilot, the effect of nudges, either positive or negative, is very small. As the pilot's population is divided in two groups, we identify different results for each group, that appear to be similar for each nudging period. Group 0 seems to have an increasing treatment effect. On the other hand, group 1 behaves differently as there is a small negative effect. Furthermore, the small sample size produces poor r-squared values for the running models with statistically insignificant results. However, a relation between energy consumption and time effects is identified, as we get statistically significant results for the specific model. Based on the described outcomes, we tried to modify our analysis by focusing on specific periods or parameters that could alter the treatment effect. Specifically, a first adjunction was to set the baseline period during January-March 2022 and compare the treatment effect during the same period in 2023. Additionally, another adjustment was to repeat the analysis by modifying the treatment group according to the days that they were exposed to nudges. However, for both analysis adjustments, the outcomes were similar to our initial analysis results, with poor r-squared values and statistically insignificant results in their majority.

Regarding the intention and motivation of PT participants to save energy, we identified that intention was slightly increased compared to the motivation to save energy. However, for both metrics, the recorded scores throughout the waves were similar without any major score difference.

8 Synthesis

We document that nudging interventions are effective in stimulating energy consumption, but only in some settings. In those cases with a positive impact, the energy savings ranged from roughly 3-4 percent for interactive nudges and reached up to 16 percent for participants who accepted an automated optimization of their consumption (in our case, optimized charging of electric vehicles). The latter is a technical feature not covered under a narrow definition of nudging. Yet, it gives valuable insights into the potential of low-interaction interventions that can relieve consumers of constant decision-making.

Nevertheless, our evidence shows that the effectiveness of nudging is inconsistent across different settings, such as countries and seasons. In two out of five pilots, we determined no significant nudging effect. In addition, in each of the five pilots, there are some groups or outcomes for which the effects were insignificant or contradictory (see Table 8-1).

In an in-depth analysis, we identify and discuss three sources of variation that constrain the effectiveness of nudging at the household level, as well as associated regulatory and market barriers. First, households have other priorities in their daily life than monitoring apps for optimizing their consumption and enforcing corresponding measures. Even though our pilots involved highly motivated households, we observed limited interaction with the apps that conveyed the nudges. This non-response results in lower average effects in the estimation. In select cases, where the kind of nudge and the app data allowed us to distinguish between responding and non-responding participants, we recognized a more pronounced nudging effect for the responding group (e.g., comparing the German and Greek pilots). However, for interactive nudges, such as feedback, and comparisons, it is challenging to distinguish between both participant groups. In those cases, it is indistinct which level of action leads to measurable behavioral change. Consequently, we recommend that service providers and policymakers choose nudges with immediate response (e.g., opt-ins). This reduces longer, convoluted chains between action and behavioral change that can decrease the effectiveness and complicate the monitoring with statistical methods.

Second, specific periods of the year, such as the Christmas holidays, appear to be a barrier to energy-saving measures. For instance, in the Belgium pilot, pupils that took the energy course immediately before the Christmas break showed no energy savings in contrast to the ones that took it afterwards. This relates to a more general point: for the nudging channel to work, attention is a pre-condition. The long operational time of our experiment allows us to compare periods with low and high attention, as well as household groups with low and high attention. This was not part of the initial set of hypotheses, but emerges as a key finding after evaluating and comparing results across all interventions included in the project.

Third, certain weather conditions lead to favorable energy consumption patterns, such that behavioral changes cannot further optimize the studied outcomes. For instance, mild winters in Belgium and Greece created little need for heating reductions. In addition, a high amount of self-generated electricity during the summer led to high autarky rates for German and Croatian prosumers – especially when combined with home battery systems. In these cases, little optimization margins exist for nudges.

Table 8-1: Overview of hypotheses and outcomes

Hypotheses		Based on	Outcome	Remark
DE1	Nudges are effective in increasing the self-consumption of participants.	DiD with sensor data (daily aggregation, baseline vs. nudge)	✓	-
DE2	Nudges are more effective in increasing the self-consumption of participants with controllable electric vehicles than of the ones without.	DiD with sensor data (peak – off-peak aggregation, baseline vs. nudge)	✓	-
DE3	Nudges are effective in reducing the overall electricity consumption of participants.	DiD with sensor data (daily aggregation)	✓	-
HR1	Nudges are effective in increasing the self-consumption of participants.	Within-subject with sensor data (1 or 2 week(s) before & during nudge)	✓	Only sufficient data for nudge 2 & 3, policy framework & weather as a confounding factor
		DiD with sensor data (daily aggregation, wash-out vs. nudge)	—	
BE1	Energy course positively impact the knowledge level of (a) children attending the energy course and of (b) their parents.	Within-subject with survey data	—	-
BE2	The energy course is effective in reducing the gas consumption of the parents.	Within-subject with sensor data (2 or 3 weeks before & after the course)	✓	Only valid for cohort 1 (cohort 2: milder winter for & courses during Christmas)
GR1	Nudges are effective in reducing participants' gas consumption.	DiD with sensor data (daily aggregation, nudge 2 vs. nudge 3)	— ✓	Only sufficient data for nudge 3, due to shorter feasible treatment period that is impacted by Christmas, and conflicting gas price increase for control group
GR2	Nudges are effective in reducing the heating time of participants.	DiD with sensor data (daily aggregation, nudge 2 vs. nudge 3)	—	

PT1	Nudges are effective in reducing participants' electricity consumption.	DiD with sensor data (daily aggregation)		-
All1	Nudges are effective in increasing participants' intention to save energy.	Within-subject with survey data	✓	Partly confirmed: DE, HR, BE, PT (not for GR)
All2	Nudges are effective in increasing the participants' motivation to save energy.	Within-subject with survey data	✓	Partly confirmed: HR, BE, PT (not for DE, GR)

Monitoring imperfect levels of interaction with apps, periods that are not suited to energy savings, or identifying limited action space due to weather conditions are important pre-conditions to effectiveness. When designing and implementing nudges, they cannot simply be viewed as confounders to a clean estimate, but as an additional lesson about real-life behavior –where there are no "clean" conditions.

As shown above, the context of everyday life can be a limitation to the effectiveness of nudging. Therefore, it is important and valuable to study the effectiveness of nudging within the context of everyday life and understand its limitations. However, other factors hindering the success of nudging exist and can be tackled. A prominent example is mismatched regulatory incentives. One example is the Croatian prosumer regulation that penalizes self-generation beyond the self-consumption level. In our pilots, it led to a (self-inflicted) curtailment of photovoltaic plants and an increase in energy consumption. By contrast, nudges designed to stimulate an increase in self-consumption were hardly of value for these participants because the regulatory system dominated their incentives. In response to this regulation, we redefined our nudges in line with the existing regulation. We developed additional features in the app to reverse the practice of curtailing to a practice of consuming more self-generated electricity.

Another exogenous event affecting the effectiveness of our nudges was the European energy crisis in 2022. It led to increased retail prices, i.e. for gas and electricity, and overall uncertainty regarding the security of supply and future framework conditions for prosumers. In our analysis, we tackled these confounding circumstances with state-of-the-art econometric approaches. When overall conditions change, the nudge effect is hard to identify as separate from bigger shifts. The Difference-in-Differences design with two levels of fixed effects is specifically suited to addressing this challenge because it only looks at the relative gap between what happens to the treatment and the control group, which have a common baseline and experience the same changes in the outside conditions. However, this evaluation strategy is demanding with respect to the data; it is "variation-hungry" in econometric language because a lot of information in the data is needed to adequately pin down the confounding factors. This can lead to constraints in statistical power problems, which we experienced in implementation.[52] The more complex the analytic strategy, the more data is needed to produce statistically significant differences. To give credible results, the estimation

[52] Power in this context is statistical power, i.e. a quantification of the chance that the model correctly rejects the null hypothesis. Power constraints are limitations to this statistical power, e.g. due to small sample sizes, and/or smaller effect size.

strategy had to be more complex to address the energy crisis, but this made it harder to obtain significant results, let alone to analyze subgroups. As smart meter data becomes more widely available, the relative costs of implementing nudging evaluations with digital tools is likely to decrease sharply. This opens up opportunities to dig deeper into open questions in the understanding of behavioral interventions. Our insights on design, implementation, and methodology gathered over three years can help future research leverage these opportunities.

One particular point that stands out in this context is that experimental designs create path dependencies in longer-term projects. We did an ex-ante power analysis, but it was based on the statistical power required for a simpler model. A recommendation for upcoming field trials is to allow a buffer on power, which can be done by adding more households, by testing fewer interventions, or by not alternating control and treatment groups. For example, the power analysis could consider a "high-parameter" scenario as a type of risk management strategy. However, all the solutions to power constraints come with either higher costs or a reduced ambition in the experiment. The balance is hard to strike. We acknowledge the limitation of our case, yet we also emphasize it as an important lesson learned. A pandemic and a global energy crisis cannot be foreseen, but there may be other projects that have a choice between two solid design options and can benefit from the considerations above.

Additionally, household-dependent characteristics that change over time are challenging to absorb – even for the proposed methods. Facing the energy crisis, we frequently collected information about changes on the household level in the surveys. A larger share of participants faced an increase in their retail prices at some point in the project. Some households responded with changes in their equipment (e.g., installing additional photovoltaic capacity). Robustness checks that exclude participants with these changes were taken as the main approach to handling this issue.

The pilots of the NUDGE project contributed substantially to the vital research field of nudges and behavioral economics. The results are in line with recent meta-studies that argue that nudging effects are context-dependent and not universally effective. While some cases are promising, nudging as a tool cannot be used as a one-size-fits-all all measure, especially when the transfer is made from the lab to the field. In the following, we reflect more specifically on existing nudging research and emphasize the importance of field trials in general, as well as specific learnings. Namely, the relevance of longitudinal data, the mix of nudges in one comparable setting, and the variety of data sources.

So far, the variety of findings in the literature makes it difficult to establish the effectiveness of nudges. Given the effectiveness of the concept, it is also difficult to identify the most promising nudges for energy savings and allocate them to households and contexts. Three methodological issues hamper this process. Firstly, methods for assessing stated preferences may not effectively capture the intuitive choices and everyday nudge effects that occur within the reality of constant decision-making. Secondly, many of the research designs of revealed preferences face severe limitations, such as insufficient sample sizes, the absence of control groups or baseline measurements, which hinder the interpretation of findings as causal effects. Studies with more extensive and diverse participant groups typically yield smaller effect sizes, which align with our findings. Lastly, the context-specific nature of nudging studies complicates comparisons between different nudge interventions. Considering these three challenges, our above-average duration of field trials allowed us to study the response of multiple nudges in one comparable setting. The control groups and high-resolution sensor data created the basis to reveal causal effects.

Based on the sensor and survey data, we can also compare the perceived and the actual impact of the nudges. Contrasting both data sources reveals drivers and barriers of behavioral change and identifies the participants' intention-action gaps. While our participants showed a high level of motivation and intention already at the start of the experiment, the intention to save energy increased during the experiment. In contrast, a decreasing trend for the motivation to save energy was recognizable for some participants. The lower motivation in combination with decreasing response rates in the survey and app can be interpreted as signs of fatigue.

More broadly, fatigue with the experiment leads to questions about ongoing learning from nudges. While the development of energy knowledge was ambiguous for the Belgium pilot, we recognized consistent, well-reflected answers from the Croatian participants about energy knowledge in a different sense. Croatian prosumers' survey responses are consistent with learning effects in the strategy towards the prosumer regulation at the beginning and the continued monitoring of compliance over time. These comparative results suggest that energy literacy is not a given outcome from nudging, but that the information provided in the nudge treatment is taken up mainly if it pertains to what participants are concerned with outside of the nudge content (as envisioned by the researcher).

Field trials enable observing intuitive decision-making as a response to nudges under real-life conditions. Laboratory experiments can only partially capture the decision environment of participants and must by design involve them in a conscious manner that can create biases in their intuitive decision-making. At the same time, we demonstrate the challenges of field trials. The points raised below are examples that relate to a broad point: the tradeoff between managing resources and striving for ambition in research. The organizational and financial expenditures only allow us to involve a limited number of participants. Still, our samples of up to 111 participants are at the upper end of the spectrum compared to other field trials. Information and communication technology are prone to outages, such as the interrupted transfer of sensor data in August 2022. Confounding factors are challenging to cover, especially when unforeseen shocks, i.e., a Europe-wide energy crisis, happen to coincide with the trial period. Still, our experiment also shows the upside of what is possible with changing circumstances: methodological knowledge becomes available to tackle these challenges, information and communication technologies are maturing, and smart meters and sensors become more broadly available as basic residential energy infrastructure. The results and experiences from NUDGE contribute to this pathway with the full execution of a large, multi-pilot field experiment.

9 Annex

9.1 Equations

9.1.1 Equations for German pilot

The main difference-in-differences regression equation is:

$$y_{it} = a_i + b_N T_{it} N_t + c G_i + d N_t + p_t + e_{it} \quad (1)$$

Where i indicates individuals and t indicates time periods (days). The indicator T equals 1 for the treated group, and zero for the non-treated group. N is a categorical variable that takes value 0 at baseline and then has 6 non-zero values. The three active nudge periods for group 1, and the three active nudge periods for group 2. The coefficient of interest is b N for all$N$ ={1;6}, which captures the DiD treatment effect from the interaction of T and N. The indicator G is a binary variable for the assignment to the two groups. Note that group assignment is stable over time, while treatment changes between group 1 and group 2. The TWFE model absorbs individual-specific intercepts (a_i) and period-specific intercepts (p_t. Robust standard errors are calculated with the common Huber-White adjustment.

For the sub-group analysis, the same regression model is adapted. G accordingly has 4 values: EV1, PV1, EV2, PV2. The marginal treatment effects for each sub-group are obtained from an additional interaction term between the treatment and the sub-group, so there are 2 coefficients b_N for each N_t.

For the deep dive in Nudge 3, we use data at hourly frequency. The regression equation is:

$$y_{it} = a_i + b_H A_{it} H_t + c H_t + p_t + e_{it} \quad (2)$$

Where A is an indicator for households that actively engage with the app, and H is a categorical variable for AM (6-10am), midday (11am-3pm), and PM (4pm-8pm). The base level is AM, and we exclude night-time hours. The coefficient of interest is b_H, which indicates whether active app users realize larger shifts during a specific Time Block H. We again use a TWFE model and robust standard errors.

9.1.2 Equations for Croatian pilot

Regression discontinuity design:

The regression discontinuity in time uses linear functions fitted to each side of the time of nudge implementation, i.e., a parametric specification with a first-degree polynomial. The formal regression equations are:

$$y_{it} = a_L + b_L p_t + e_{it} \quad \text{with} \quad -\mathrm{h} \leq p_t < 0 \quad (3)$$

$$y_{it} = a_R + b_R p_t + e_{it} \quad \text{with} \quad 0 \leq p_t < +\mathrm{h} \quad (4)$$

where i indicates individuals and t indicates time periods (days). The first equation represents the time period before implementation ("left" of 0 on timeline: L), and the second equation represents the timeline after implementation ("right" of 0 on timeline: R). The main model uses a window of 7 days before the nudge starts and the first 7 days of the intervention (bandwidth h=7). We stack the two groups due to power constraints, so there is a single time indicator centered to 0, i.e. t={−7;7}. bL and bR represent the linear time trends before and after the nudge becomes effective. The resulting estimate of the treatment effect is aR – aL , i.e. the discontinuous break at the time of treatment. The estimate indicates the change in the outcome (autarky or self-consumption) when the nudge is turned on. In practice, implementation is done with the RD-robust package by Calonico et al. (2017)[53]

The same specification is used in robustness tests for a different bandwidth h (i.e., longer time period around implementation). Testing for weather as a potential confounder is done by using the same equations with radiation as the outcome variable.

The short-term difference-in-differences model is:

$$y_{it} = a_i + b_N\, T_i\, N_t + c\, G_i + d\, N_t + p_t + e_{it} \qquad (5)$$

Where i indicates individuals and t indicates time periods (days) using a window of 7 days before the nudge starts (common to both groups), and the first 7 days of the intervention for the respective group (staggered treatment). The indicator T equals 1 for the treated group, and zero for the non-treated group. N is a categorical variable that takes value 0 before the intervention starts and 1 afterwards. The model is evaluated separately for each group and nudge period. As in the German case, the TWFE model absorbs individual-specific intercepts (ai) and period-specific intercepts (pt), and reports Huber-White adjusted standard errors.

9.1.3 Equations for Belgian pilot

For the Belgian pilot sensor data analysis, we applied two variations of random intercept models in all cases following the same logic. In the first, model(x) and model(x)b series, kWh (kwh) and kwh with outliers removed (kwh_o) are used as our outcome/dependent variable. Outliers determined by multiplying the upper interquartile range (IQR) by 5. For these models, we include heating degree days, base temperature of 23: hdd23 as our control variable. In model(x)c and model(x)d, usage is normalized through inclusion of hdd23 in our daily consumption:

$$\text{nkWh} = \text{kWhNA} / \text{hdd23} \qquad (6)$$

model(x)d is a variation nkWh, but again with outliers removed by multiplying the upper IQR by 5.

[53] For further information please visit: https://journals.sagepub.com/doi/10.1177/1536867X1701700208

Finally, for each of these four variations, we add the 14 day period before and after the intervention (real_int_week14), or the 21 day period before and after the intervention (real_int_week21) as categorical variable. This procedure is repeated for both cohorts.

Random intercept models equation is as follow:

$$Y = X\beta + Zu + \varepsilon \tag{7}$$

where

Y is the vector of observed energy consumption values for each household, either normalized, or with outliers removed

X is the matrix of design variables, including an intercept, the intervention period and hdd in the case where we don't use normalized data.

β is the vector of fixed effect coefficients, representing the average effect of each independent variable on energy consumption

Z is the matrix of random effect variables, in our case the nudge_id

u is the vector of random effect terms, representing the deviation of each household's energy consumption from the overall mean

ε is the vector of residual errors, representing the variation in energy consumption that is not explained by the independent variables or random effects.

As such for model(x)a and model(x)b where we use non-normalized consumption and control for temperature using hdd23, we have the following equation:

$$\mathrm{kWh}_i = \beta_0 + \beta_1(\mathrm{real\_int\_week14}) + \beta_2(\mathrm{hdd23}) + b_{\mathrm{nudgeid}j} + \epsilon_i \tag{8}$$

and

$$\mathrm{kWh\_o}_i = \beta_0 + \beta_1(\mathrm{real\_int\_week14}) + \beta_2(\mathrm{hdd23}) + b_{\mathrm{nudgeid}j} + \epsilon_i \tag{9}$$

For model(x)c and model(x)d, where we used normalized consumption, we have the following equation, first with outliers included and second with outliers removed:

$$\mathrm{nkWh}_i = \beta_0 + \beta_1(\mathrm{real\_int\_week14}) + b_{\mathrm{nudgeid}j} + \epsilon_i \tag{10}$$

and

$$\mathrm{nkWh\_o}_i = \beta_0 + \beta_1(\mathrm{real\_int\_week14}) + b_{\mathrm{nudgeid}j} + \epsilon_i \tag{11}$$

For our 21day variation, our equation correspondingly changes, with model(x)a and model(x)b having the following equations:

$$\mathrm{kWh}_i = \beta_0 + \beta_1(\mathrm{real\_int\_week21}) + \beta_2(\mathrm{hdd23}) + b_{\mathrm{nudgeid}j} + \epsilon_i \tag{12}$$

and

$$kWh\_o_i = \beta_0 + \beta_1(real\_int\_week21) + \beta_2(hdd23) + b_{nudgeid_j} + \epsilon_i \quad (13)$$

Finally for model(x)c and model(x)d, where we used normalized consumption, we thus have the following equations:

$$nkWh_i = \beta_0 + \beta_1(real\_int\_week21) + b_{nudgeid_j} + \epsilon_i \quad (14)$$

and

$$nkWh\_o_i = \beta_0 + \beta_1(real\_int\_week21) + b_{nudgeid_j} + \epsilon_i \quad (15)$$

9.1.4 Equations for Greek pilot

For the analysis of the nudge 3 effect in the Greek pilot, we applied the DiD model twice, once for each KPI, gas consumption and heating time.

Regarding gas consumption, we worked with the following baseline model:

$$LnEnergyIn_{ij} = \beta_0 + \beta_1 \cdot group_i + \beta_2 \cdot intervention_j + \beta_3 \cdot group_i \cdot intervention_j + \varepsilon_{ij} \quad (16)$$

where *LnEnergyIn* represents the logarithmic transformation of gas consumption in KWh; *group* is an indicator variable denoting the user's participating group, which equals 0 for the control and 1 for the intervention groups, respectively; *intervention* is an indicator variable, with 0 for the pre-treatment and 1 for the treatment period, and the *group*intervention* interaction term is the product of the two indicator variables. The coefficient β_3 of this interaction term yields the treatment effect. Since the dependent variable is logarithmically transformed, the estimated effect corresponds to the % change in the value of the EnergyIn.

For the analysis of the heating time, the basic DiD estimator model is:

$$BoilerHeat_{ij} = c_0 + c_1 \cdot group_i + c_2 \cdot intervention_j + c_3 \cdot group_i \cdot intervention_j + c_4 \cdot boilerT_i + \varepsilon_{ij}$$

where we used the same independent variables as for the previous model, except that $BoilerHeat$ represents the heating time and $boilerT$ is a binary variable denoting the type of boiler that each household has (0 for OpenTherm and 1 for OnOff boilers).

9.1.5 Equations for Portuguese pilot

For the provided analysis, we worked with the following baseline model:

$$LnTotalConsumption_{ij} = \beta_0 + \beta_1 \cdot group_i + \beta_2 \cdot intervention_j + \beta_3 \cdot group_i \cdot intervention_j + \varepsilon_{ij}$$

where $LnTotalConsumption$ represents the logarithmic transformation of electricity consumption in Wh; $group$ is an indicator variable denoting the user's participating group, which equals 0 for the control and 1 for the intervention groups, respectively; $intervention$ is an indicator variable, with 0 for the pre-treatment and 1 for the treatment period, and the $group \cdot intervention$ interaction term is the product of the two indicator variables. The coefficient β_3 of this interaction term yields the treatment effect. Since the dependent variable is logarithmically transformed, the estimated effect corresponds to the % change in the value of the EnergyIn.

9.2 Post-intervention

9.2.1 German pilot

For the post-intervention analysis, we compare the baseline period in 2022 with the post-intervention period in 2023. Given the potential impact of the energy crisis and the seasonal differences, the simple development of each group over time would not be indicative of lasting nudge effects. Therefore, the analysis again uses the difference-in-differences design, with group 1 coded as the "treated group". Hence, the results reveal whether the two groups developed differently from before the experiment to the conclusion of the experiment. The coefficient for the post-estimation period indicates whether the outcome changed more for group 1 relative to the development for group 2. Noting that group 1 received all three nudges first over the course of the experiment, group 1 is already several weeks removed from the final nudge at the start of the post-intervention phase, while it is "fresh" for group 2. The results are shown in TableAnnex 1 for the same three outcomes as the main analysis: autarky, self-consumption, and household consumption.

TableAnnex 1: *Post-Intervention Results for German pilot*

	(1) Autarky	(2) Self-Consumption	(3) Household Consumption
Post-Intervention	-0.00285	0.0580^{***}	0.00789
	(-0.82)	(4.08)	(0.60)
Constant	0.557^{***}	5.810^{***}	6.311^{***}
	(501.74)	(1338.92)	(1586.33)
R2	0.779	0.752	0.593
Observations	17554	17471	16868

Notes: *Difference-in-differences estimation for the comparison between baseline and post-intervention phases for dependent variables autarky (0 to 1 ratio), self-consumption, and household consumption (log-transformed). Household consumption is total consumption excluding EV charging. Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

For autarky, we find no post-intervention effects. The coefficient is very close to zero and not statistically significant. For self-consumption, there is a significant post-intervention increase of 5.8%. This indicates that group 1 increased self-consumption more than group 2 relative to the baseline. Given that group 1 is further removed from the large nudge in time, this result does not conform to expectations. For household consumption there is again no significant effect – the coefficient is close to zero and not significantly different from zero at any conventional level of significance.

In light of the findings of the main analysis, we looked into the sub-groups to understand whether the self-consumption effect could be driven by long-run adaptations in charging behavior in the EV group. However, this explorative analysis suggests that it is the PV 1 group that drives the average increase. We cannot substantiate the underlying mechanism further, but the final survey wave does indicate that several households made changes by investing in new assets or expanding their PV capacity in 2023. Such factors would not be directly attributable to the nudging, although the project may trigger related investments. With full exposure to higher energy prices for households occurring only at the turn of the year, and considering subsequent changes to regulatory conditions throughout 2023, we interpret the results with caution. Overall, there is no strong evidence for substantial group differences between the baseline and post-intervention periods.

9.2.2 Croatian pilot

In the Croatian pilot, there is no solid baseline against which to evaluate the post-intervention behavior. However, the design with the switch in the treatment group allows us to draw insights on whether behavioral interventions last over time. To this end, we compare the development between nudge 3 and the post intervention, again using differences-in-differences. The idea is that group 1 gets treated first, then the nudge is taken away for several weeks (while group 2 is treated), before both groups are again equal in the post-intervention. Hence, the estimation compares only the first phase of Nudge 3 and the post intervention. The difference-in-differences effect captures the effect of *taking away* the nudge earlier. If it does matter how long participants are removed from the last treatment, we would have negative effects for group 1: group 2 would catch up in the meantime, while the effects fade for group 1. TableAnnex 1 shows the results from this exercise.

There is no significant effect on autarky, which indicates that the two groups did not diverge between the third intervention and the post-intervention periods. For both self-consumption and total consumption, the effect is negative. The results indicate that group 1 has dropped self-consumption more strongly than group 2, which is consistent with a fading effect. However, total consumption dropped by about the same order of magnitude, which would indicate the opposite: group 1 having a more lasting effect. Overall, their results provide no clear evidence that the timing of the nudge matters. Importantly, a causal interpretation requires the assumption that the two groups respond equally strongly to the nudge itself, and the timing of being treated first is the only difference. In our context, this is a strong assumption, as the treatment effects have shown to be time-dependent. Without more baseline data, it is not feasible to assert fading or learning effects in a statistically credible manner.

TableAnnex 2: Croatian post intervention text

	(3) Autarky	(6) Self-Consumption	(9) Total Consumption
Treatment Effect	0.00760	-0.204***	-0.212***
	(1.09)	(-5.36)	(-5.85)
Constant	0.620***	5.644***	6.305***
	(188.04)	(329.87)	(382.69)

R2	0.609	0.634	0.615
Observations	11217	10716	10837

Notes: *Difference-in-differences estimation for the comparison between nudge 3 and post-intervention phases for dependent variables autarky (0 to 1 ratio), self-consumption, and total consumption (log-transformed). Models include time and household fixed effects. Robust standard errors (Huber-White) in parentheses. Significance levels:* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

9.2.3 Belgian pilot

Given the start of the intervention for cohort 1 in January 2022, we are able to do a year-on-year analysis to see whether we find differences in energy consumption for our cohort 1 participants, one year later. Our results show that, compared to the 21 days before the intervention, the energy consumption during a 21-day period one year later (i.e.: 365 + 21), is statistically significantly lower. Our results hold true in all conditions, both when considering normalized consumption and when controlling for the weather through HDD see models 5a, 5b, 5c and 5d).

TableAnnex 3: Random intercept model results for cohort 1 one year before and after intervention

	Dependent variable:			
	kWh	kWh (no outliers)	normalized kWh	normalized kWh (no outliers)
	Model 5a	Model 5b	Model 5c	Model 5d
intervention period, 21 before and 365 days after (1)	-76.65**	-74.75**	-4.73**	-4.63**
HDD (23°C base)	5.14**	4.82**		
Constant	64.30**	66.00**	8.89**	8.77**
Observations	1,019	1,005	1,019	1,013
Log Likelihood	-5,513.23	-5,343.88	-2,660.85	-2,569.29
Akaike Inf. Crit.	11,036.46	10,697.76	5,329.70	5,146.59
Bayesian Inf. Crit.	11,061.10	10,722.33	5,349.41	5,166.27
Note:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

9.2.4 Greek pilot

In the case of the Greek pilot, the start of the post-intervention period is in April '23, when the heating demand and gas consumption in Greece are minimal. This precludes any reliable estimation of nudging effect on the energy behavior for heating purposes. Meaningful gas consumption levels in the post-intervention period would be obtained by December '23, about 8 months away from the end of the third intervention period and beyond the duration of the NUDGE project.

Even if this roadblock could be overcome, it would still be difficult to analyze the long-lasting effects of nudging since the nudging features remain part of the app and are not removed once they are added to it. This is an integral part of the domX customer policy for apparent commercial reasons. Overall, it was not possible to conduct any reasonable analysis of the nudging long-lasting effects in the case of the Greek pilot.

9.2.5 Portuguese pilot

For the post-intervention analysis, we compare the electricity consumption between a two-month baseline period, during which the nudge.it app was launched (April-May '22) with the same months in the post-intervention period in (April-May '23). We carry out a difference-in-differences analysis focusing on one of the two groups of pilot participants, the one that was first exposed to each nudging intervention. The coefficient of treatment derived from the DiD analysis denotes the relative difference in the consumption during the post-intervention period between the two groups, the one that was more recently exposed to the last nudge against the one that did so a month ago. This way, we explore how much the recency of nudging intervention may make a difference in the electricity consumption. Table 9-1 reports the coefficient values under the basic DiD models and its variants accounting for household only, time only and combined household and time fixed effects.

Table 9-1: Difference-in-differences estimation for the comparison between baseline and post-intervention phases for electricity consumption in Wh (log-transformed). Models include time and household fixed effects.

Model	Coefficient of treatment effect	p-value	R^2 (between)
Basic	0.0151	0.488	0.0130
+household FE	0.0174	0.7945	0.105
+ time FE	0.0174	0.2604	0.0074
+household and time FE	0.0179	0.7902	0.0001

The provided results in Table 9-1 indicate a minimal effect with an increase of electricity consumption during the post-intervention period. However, the outcomes are statistically insignificant and with small R^2 values.

9.3 Additional evaluations

9.3.1 Logarithmic transformation for the Belgian pilot

As our boxplot suggests in Figure 5-3, we have several outliers in our data. While we considered the impact of outliers by removing data that falls outside the five times IQR range, it is possible that this was too conservative. One solution to account for this, is by log transforming our outcome variable. This can reduce the influence of extreme values, making the distribution more symmetric and potentially bringing the outliers closer to the center of the data.

TableAnnex 4: Logged kWh consumption results for cohort 2 and cohort 2, 3 week variation

	Dependent variable:

	Cohort 1	Cohort 2	Cohort 1	Cohort 2
	log(kwh)	log(kwh)	log(normalized kWh)	log(normalized kWh)
	Model 6a	Model 6b	Model 6c	Model 6d
intervention period, 21 days (1)	-0.42**	0.30**	-0.36**	0.35**
HDD (23°C base)	-0.01	0.07**		
Constant	5.48**	3.10**	2.46**	1.50**
Observations	608	939	608	939
Log Likelihood	-546.58	-652.63	-565.34	-633.8
Akaike Inf. Crit.	1,103.16	1,315.26	1,138.69	1,275.60
Bayesian Inf. Crit.	1,125.21	1,339.48	1,156.33	1,294.98
Note:	1: pre-intervention period as reference; * p<0.05, ** p<0.01			

As seen in TableAnnex 5, our results remain consistent when log transforming our outcome variable: we see statistically significant reductions in cohort 1, while cohort 2 sees statistically significant increases.

Recall also, that in order to avoid artificially low means, we transformed all 0 values for kWh to NAs (missing data). To assess whether this had an impact on our overall analysis, we repeated our random intercept modeling, for both cohorts using the original kWh values. Results align with our initial analysis, with both 14 and 21 day variations showing decreased consumption in cohort 1 (model 7a and model 7b), while we see increased consumption in cohort 2 (model 7c and model 7d). A further analysis, not shown, with an extremely conservative IQR cut-off of 1, as opposed to the more commonly applied 5, has similarly had the same general results, with our hypothesis being supported for cohort 1, but not for cohort 2.

TableAnnex 6: Random intercept model with original kWh values, 14 and 21 day variations, cohort 1 and cohort 2

	Dependent variable:			
	Cohort 1	Cohort 1	Cohort 2	Cohort 2
	kWh	kWh	kWh	kWh
	Model 7a	Model 7b	Model 7c	Model 7d
intervention period, 14 days (1)	-53.26**		40.12**	
intervention period, 21 days (1)		-73.57**		42.81**
HDD (23°C base)	-0.72	-1.62	6.51**	7.26**

Constant	243.26**	268.07**	-17.5	-32.03*
Observations	410	611	632	947
Log Likelihood	-2,374.59	-3,536.66	-3,376.48	-5,097.28
Akaike Inf. Crit.	4,759.17	7,083.32	6,762.96	10,204.57
Bayesian Inf. Crit.	4,779.25	7,105.40	6,785.20	10,228.83
Note:	1: pre-intervention period as reference; * $p<0.05$, ** $p<0.01$			